Mushroom classification EDA VSCode Notebook

In [ ]:
```
# Imports
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use("ggplot")
```

In [ ]:
```
# Read data into pandas Dataframe
data_location = "../data/mushrooms.csv"
mushrooms = pd.read_csv(data_location)
mushrooms.head(5)
```

Out[ ]:

	class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	ring-number	ring typ
0	p	x	s	n	t	p	f	c	n	k	...	s	w	w	p	w	o	
1	e	x	s	y	t	a	f	c	b	k	...	s	w	w	p	w	o	
2	e	b	s	w	t	l	f	c	b	n	...	s	w	w	p	w	o	
3	p	x	y	w	t	p	f	c	n	n	...	s	w	w	p	w	o	
4	e	x	s	g	f	n	f	w	b	k	...	s	w	w	p	w	o	

5 rows × 23 columns

In [ ]:
```
# Basic dataframe analysis
mushrooms.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8124 entries, 0 to 8123
Data columns (total 23 columns):
 #   Column                    Non-Null Count  Dtype 
---  ------                    --------------  ----- 
 0   class                     8124 non-null   object
 1   cap-shape                 8124 non-null   object
 2   cap-surface               8124 non-null   object
 3   cap-color                 8124 non-null   object
 4   bruises                   8124 non-null   object
 5   odor                      8124 non-null   object
 6   gill-attachment           8124 non-null   object
 7   gill-spacing              8124 non-null   object
 8   gill-size                 8124 non-null   object
 9   gill-color                8124 non-null   object
 10  stalk-shape               8124 non-null   object
 11  stalk-root                8124 non-null   object
 12  stalk-surface-above-ring  8124 non-null   object
 13  stalk-surface-below-ring  8124 non-null   object
 14  stalk-color-above-ring    8124 non-null   object
 15  stalk-color-below-ring    8124 non-null   object
 16  veil-type                 8124 non-null   object
 17  veil-color                8124 non-null   object
 18  ring-number               8124 non-null   object
 19  ring-type                 8124 non-null   object
 20  spore-print-color         8124 non-null   object
 21  population                8124 non-null   object
 22  habitat                   8124 non-null   object
dtypes: object(23)
memory usage: 1.4+ MB
```

No null values, all features are categorical.

In [ ]:
```
# Replace feature values with their full names
# mappings written manually based on Kaggle data explorer
feature_value_mappings = {
    "class": {"e": "edible", "p": "poisonous"},
    "cap-shape": {"b": "bell", "c": "conical", "x": "convex", "f": "flat", "k": "knobbed", "s": "sunken"},
    "cap-surface": {"f": "fibrous", "g": "grooves", "y": "scaly", "s": "smooth"},
    "cap-color": {"n": "brown", "b": "buff", "c": "cinnamon", "g": "gray", "r": "green", "p": "pink", "u": "purple", "e": "red
```

```
    "bruises": {"t": "bruises", "f": "no"},
    "odor": {"a": "almond", "l": "anise", "c": "creosote", "y": "fishy", "f": "foul", "m": "musty", "n": "none", "p": "pungent
    "gill-attachment": {"a": "attached", "d": "descending", "f": "free", "n": "notched"},
    "gill-spacing": {"c": "close", "w": "crowded", "d": "distant"},
    "gill-size": {"b": "broad", "n": "narrow"},
    "gill-color": {"k": "black", "n": "brown", "b": "buff", "h": "chocolate", "g": "gray", "r": "green", "o": "orange", "p": "
    "stalk-shape": {"e": "enlarging", "t": "tapering"},
    "stalk-root": {"b": "bulbous", "c": "club", "u": "cup", "e": "equal", "z": "rhizomorphs", "r": "rooted", "?": "missing"},
    "stalk-surface-above-ring": {"f": "fibrous", "y": "scaly", "k": "silky", "s": "smooth"},
    "stalk-surface-below-ring": {"f": "fibrous", "y": "scaly", "k": "silky", "s": "smooth"},
    "stalk-color-above-ring": {"n": "brown", "b": "buff", "c": "cinnamon", "g": "gray", "o": "orange", "p": "pink", "e": "red"
    "stalk-color-below-ring": {"n": "brown", "b": "buff", "c": "cinnamon", "g": "gray", "o": "orange", "p": "pink", "e": "red"
    "veil-type": {"p": "partial", "u": "universal"},
    "veil-color": {"n": "brown", "o": "orange", "w": "white", "y": "yellow"},
    "ring-number": {"n": "none", "o": "one", "t": "two"},
    "ring-type": {"c": "cobwebby", "e": "evanescent", "f": "flaring", "l": "large", "n": "none", "p": "pendant", "s": "sheathi
    "spore-print-color": {"k": "black", "n": "brown", "b": "buff", "h": "chocolate", "r": "green", "o": "orange", "u": "purple
    "population": {"a": "abundant", "c": "clustered", "n": "numerous", "s": "scattered", "v": "several", "y": "solitary"},
    "habitat": {"g": "grasses", "l": "leaves", "m": "meadows", "p": "paths", "u": "urban", "w": "waste", "d": "woods"}
}
for column, mapping in feature_value_mappings.items():
    if column in mushrooms.columns:
        mushrooms[column] = mushrooms[column].replace(mapping)
    else:
        print(f"Column {column} not found in mushrooms.csv")
mushrooms.head(5)
```

Out[ ]:

	class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color
0	poisonous	convex	smooth	brown	bruises	pungent	free	close	narrow	black	...	smooth	white	white	partial	white
1	edible	convex	smooth	yellow	bruises	almond	free	close	broad	black	...	smooth	white	white	partial	white
2	edible	bell	smooth	white	bruises	anise	free	close	broad	brown	...	smooth	white	white	partial	white
3	poisonous	convex	scaly	white	bruises	pungent	free	close	narrow	brown	...	smooth	white	white	partial	white
4	edible	convex	smooth	gray	no	none	free	crowded	broad	black	...	smooth	white	white	partial	white

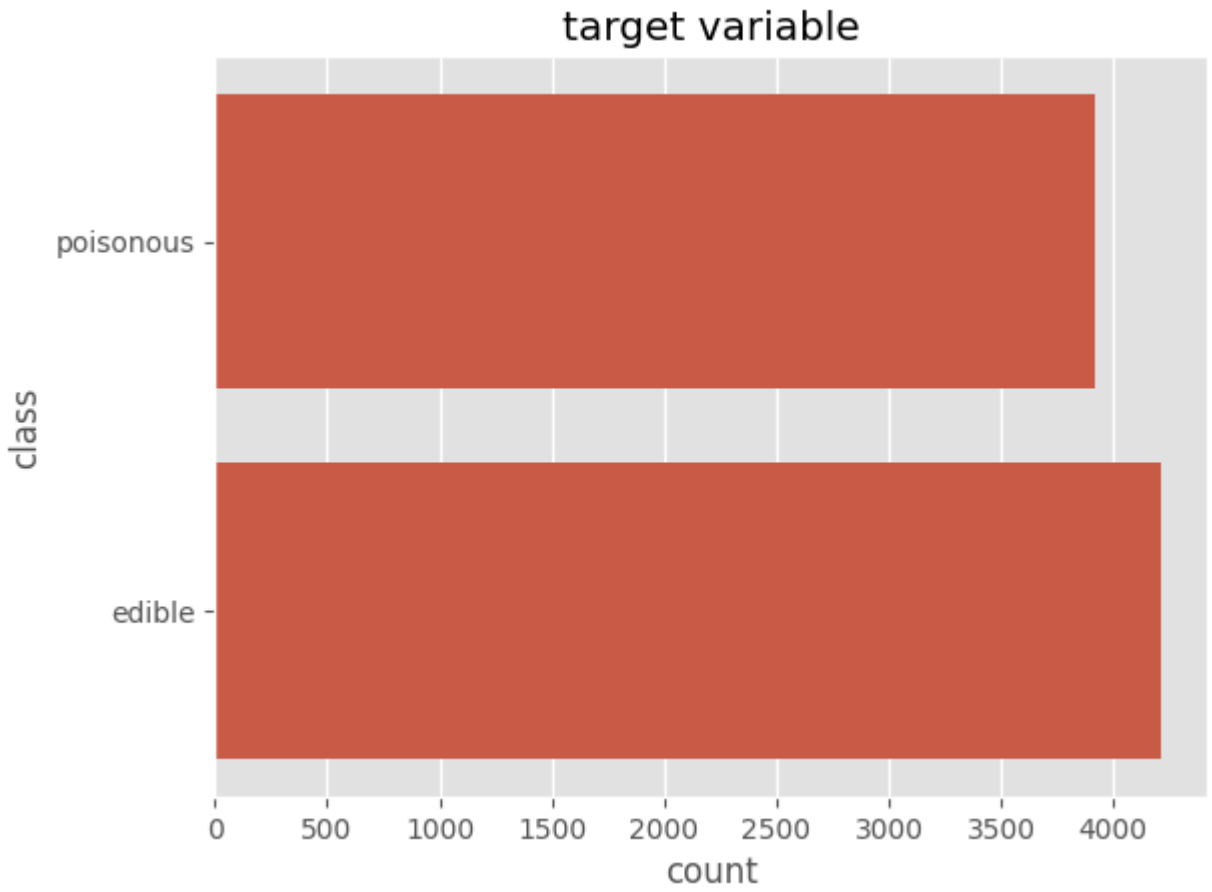
5 rows × 23 columns

In [ ]:
```
# Target variable analysis
print(mushrooms["class"].value_counts())
print()
print(mushrooms["class"].value_counts() / len(mushrooms))
sns.countplot(mushrooms["class"])
plt.title("target variable")
```

```
class
edible       4208
poisonous    3916
Name: count, dtype: int64

class
edible       0.517971
poisonous    0.482029
Name: count, dtype: float64
```

Out[ ]: Text(0.5, 1.0, 'target variable')



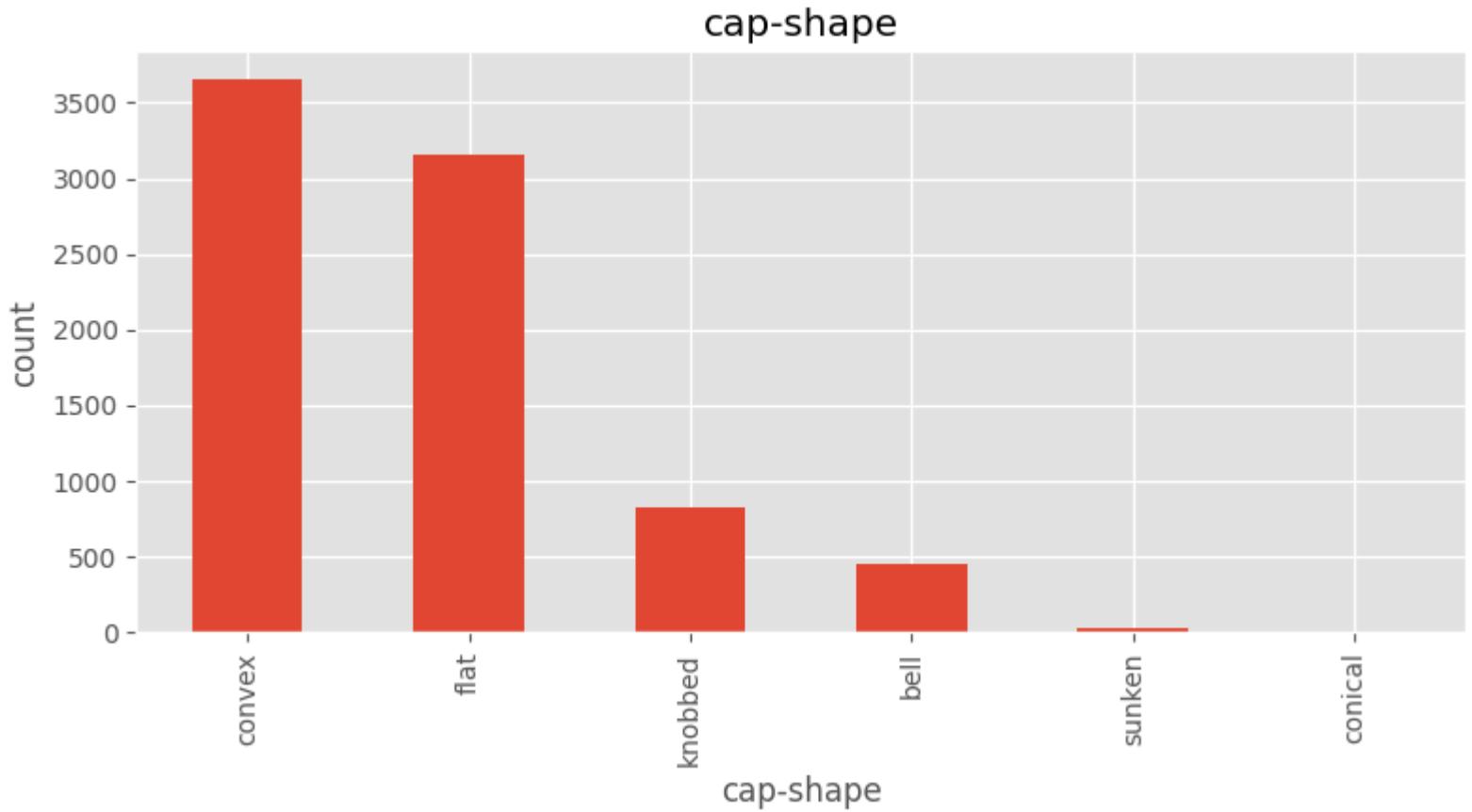
Target variable analysis

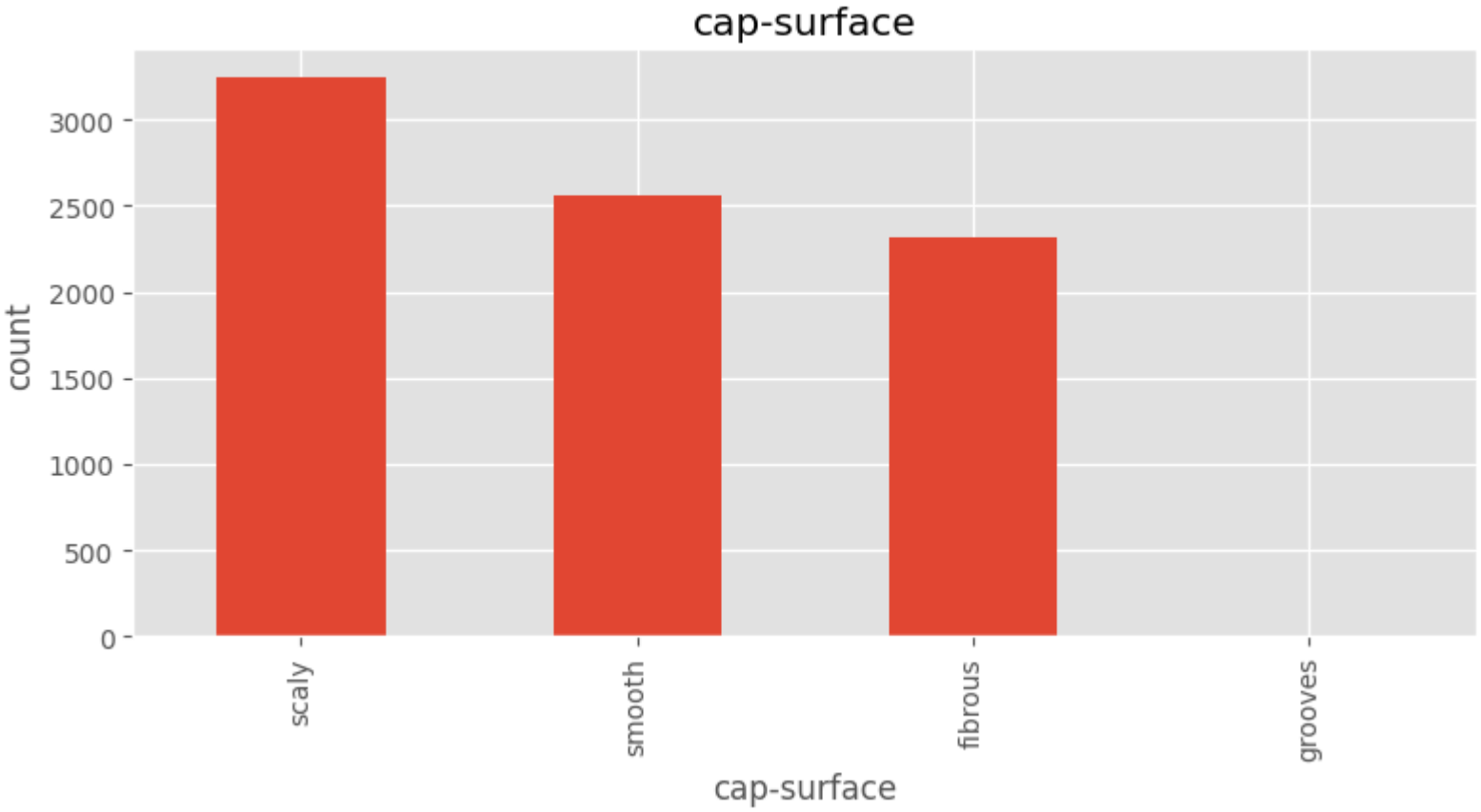
Fairly balanced target variable (51% edible 48% poisonous).

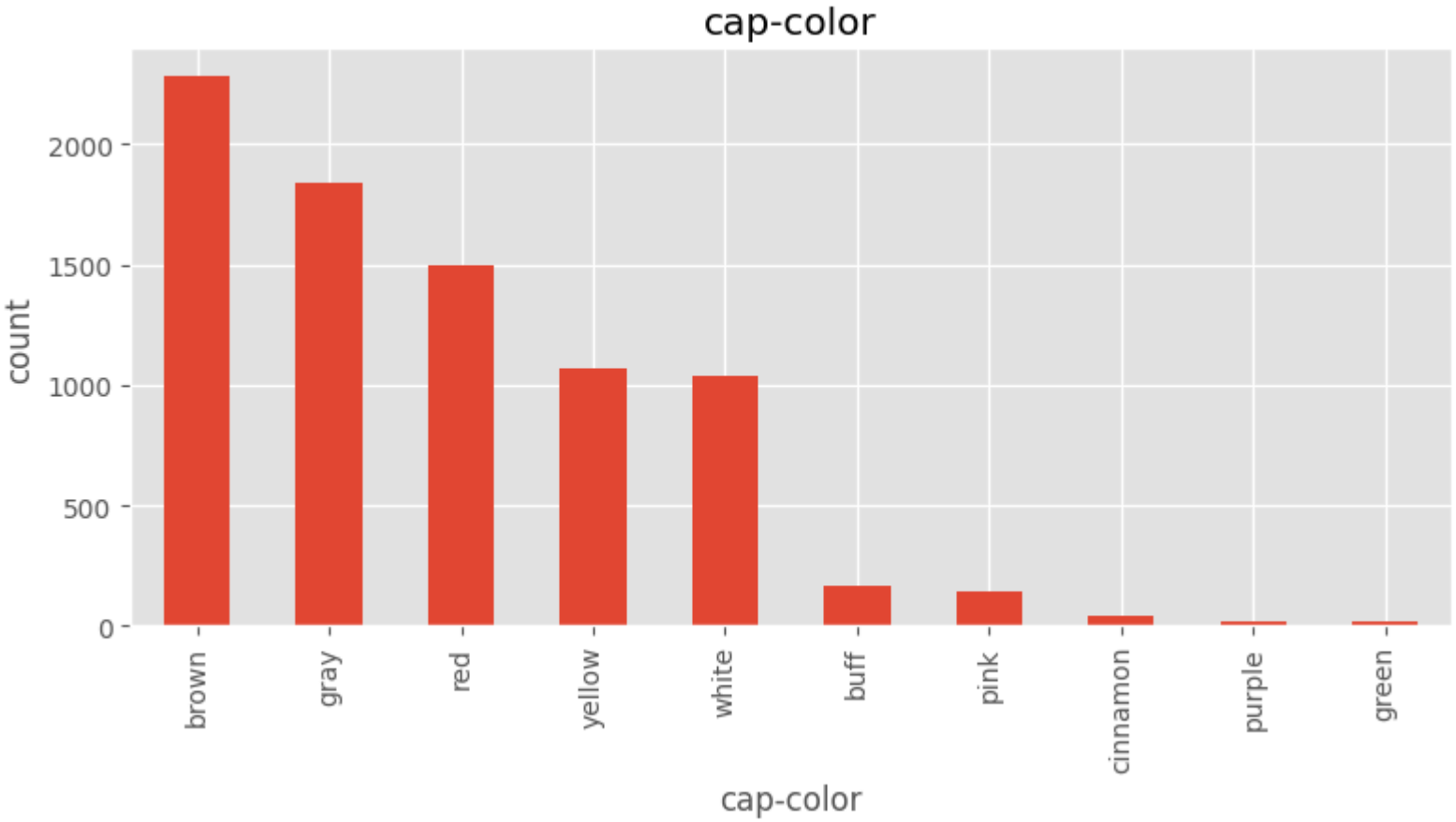
In [ ]:
```
# All features univariate analysis
for column in mushrooms:
    if column == "class":
        continue
    plt.figure(figsize=(20,4))
    plt.subplot(121)
```

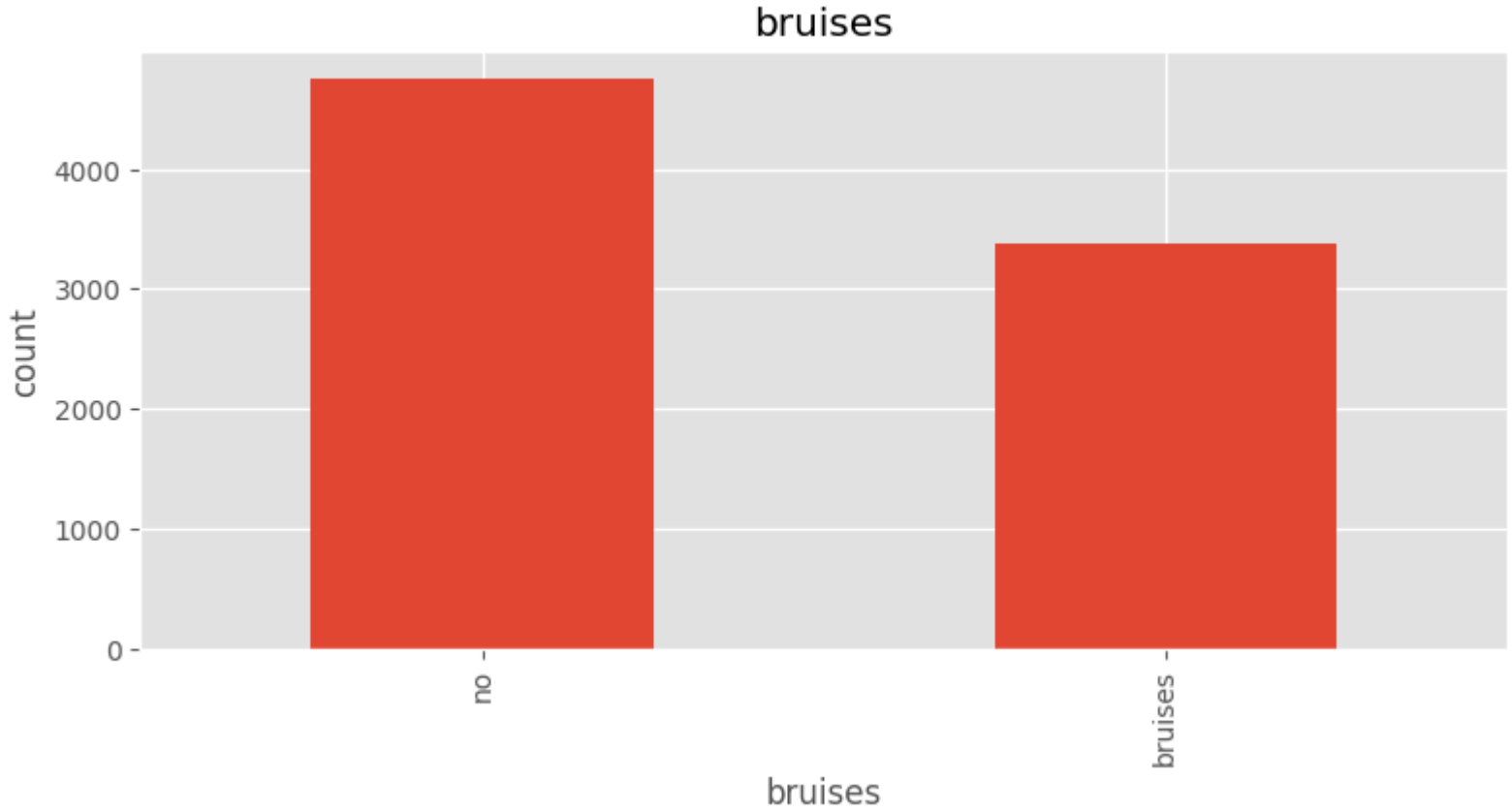
```
    mushrooms[column].value_counts().plot(kind="bar")
    plt.xlabel(column)
    plt.ylabel("count")
    plt.title(column)
```

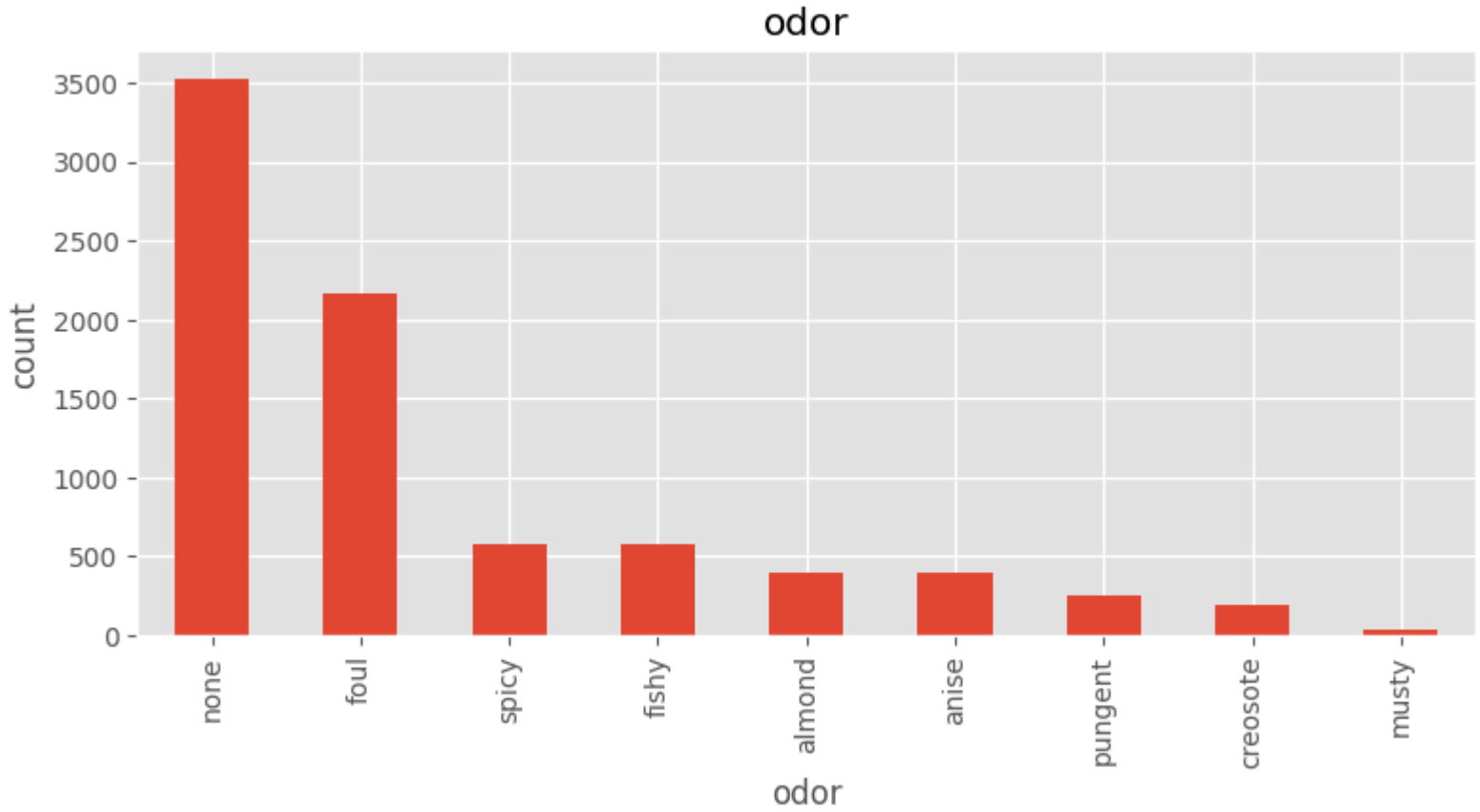
```
<ipython-input-114-104510d4860f>:6: RuntimeWarning: More than 20 figures have been opened. Figures created through the pyplot i
nterface (`matplotlib.pyplot.figure`) are retained until explicitly closed and may consume too much memory. (To control this wa
rning, see the rcParam `figure.max_open_warning`). Consider using `matplotlib.pyplot.close()`.
  plt.figure(figsize=(20,4))
```

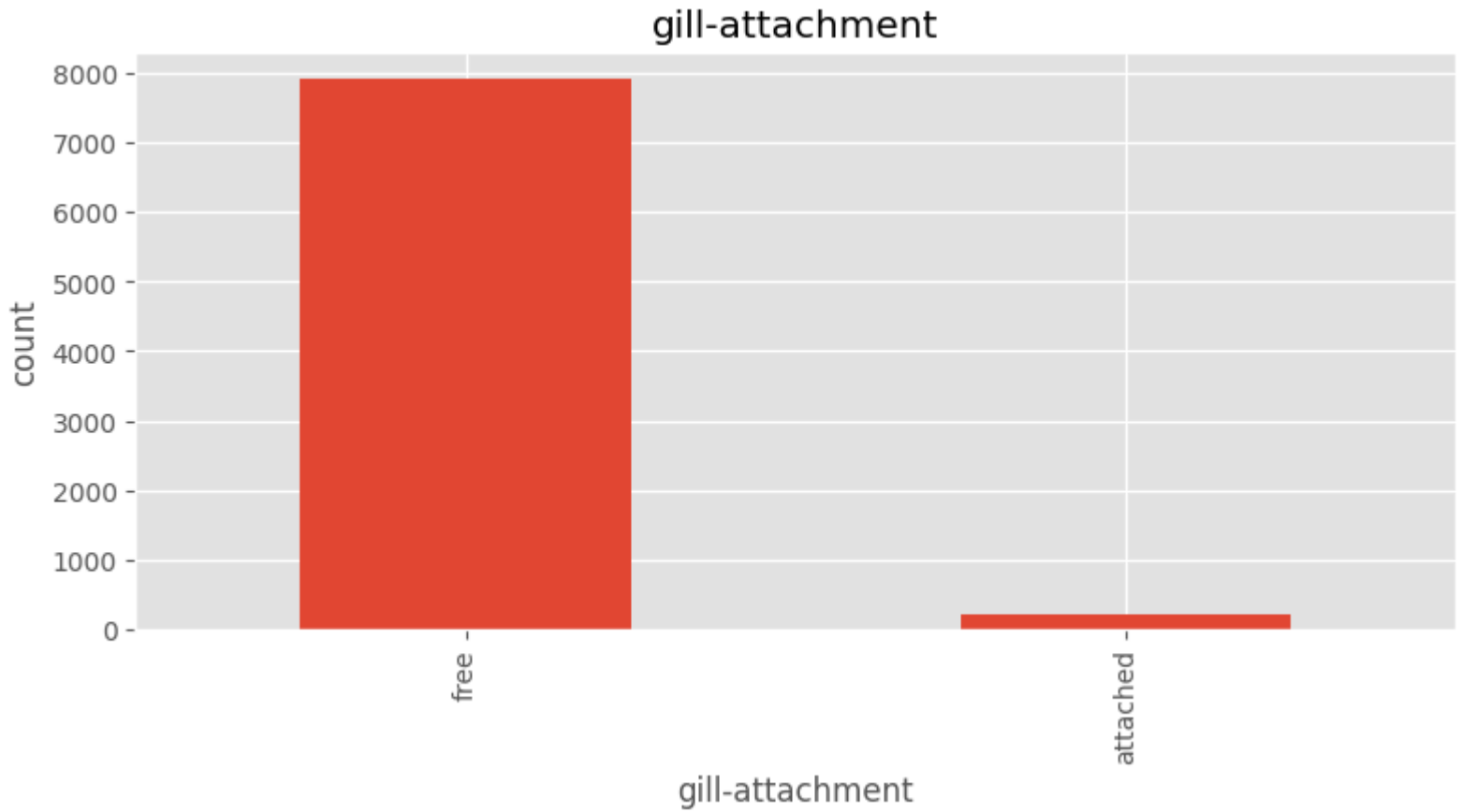


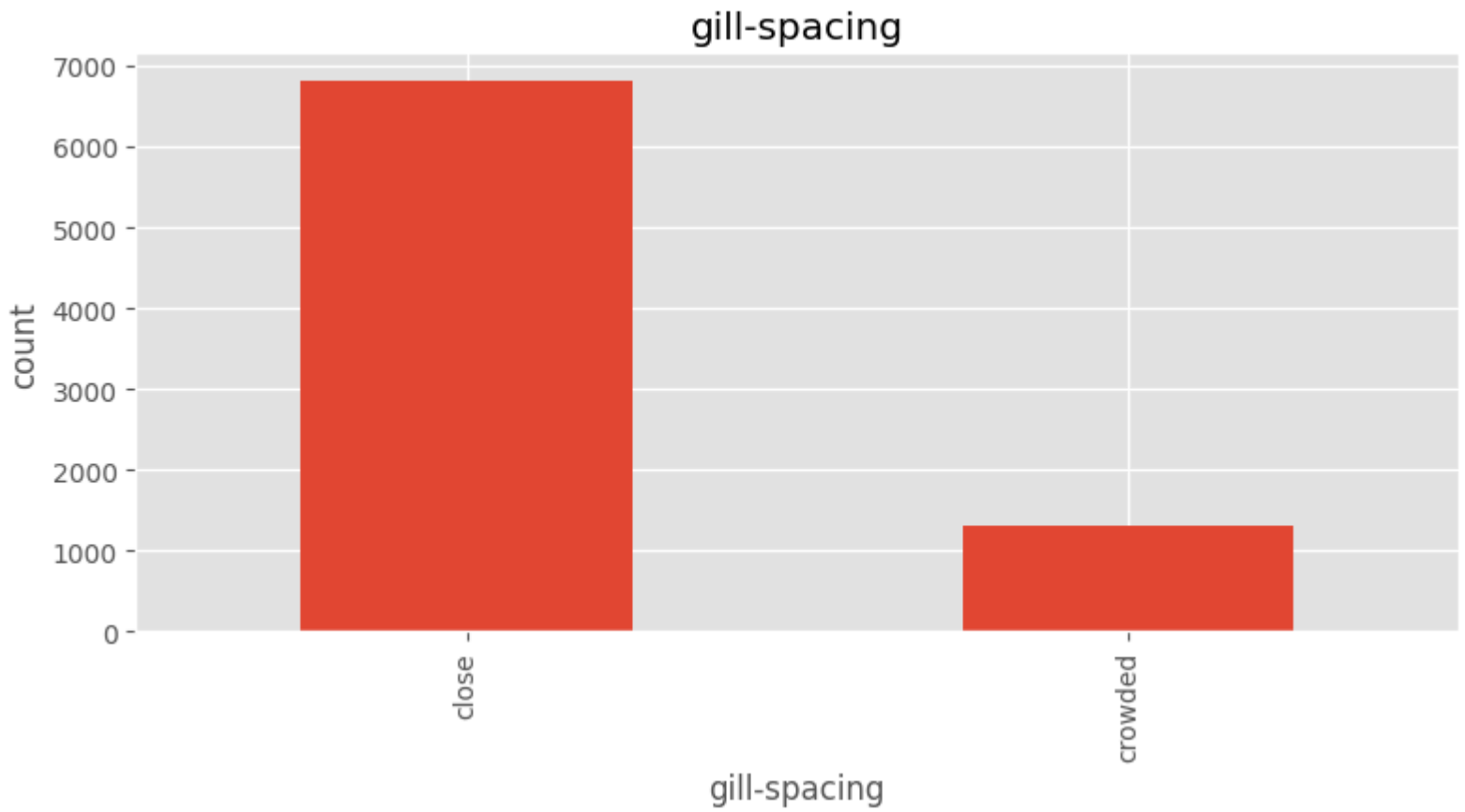

cap-surface
count
3000
2500
2000
1500
1000
500
0
scaly
smooth
fibrous
grooves
cap-surface

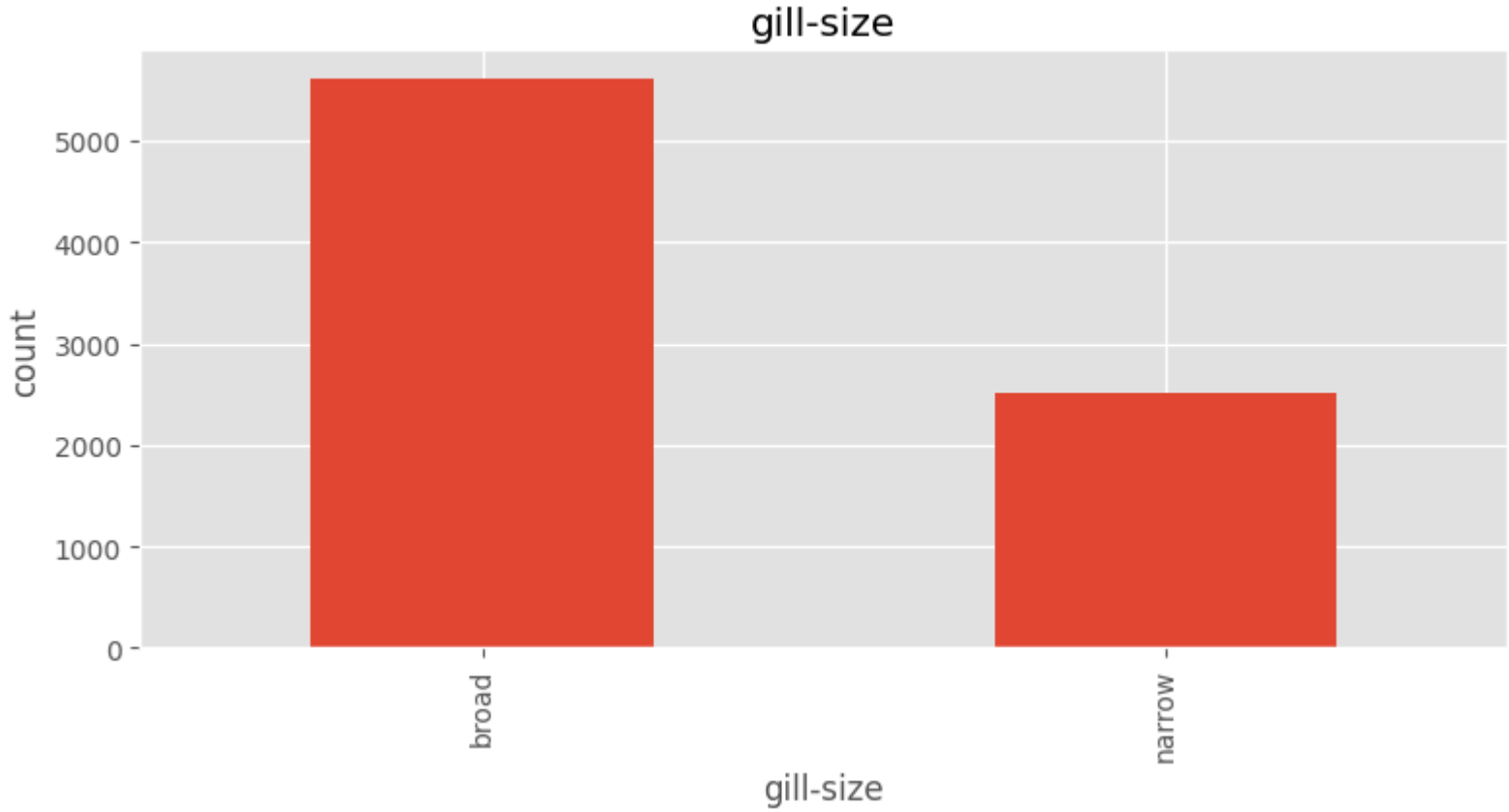

cap-color
count
2000
1500
1000
500
0
brown
gray
red
yellow
white
buff
pink
cinnamon
purple
green
cap-color

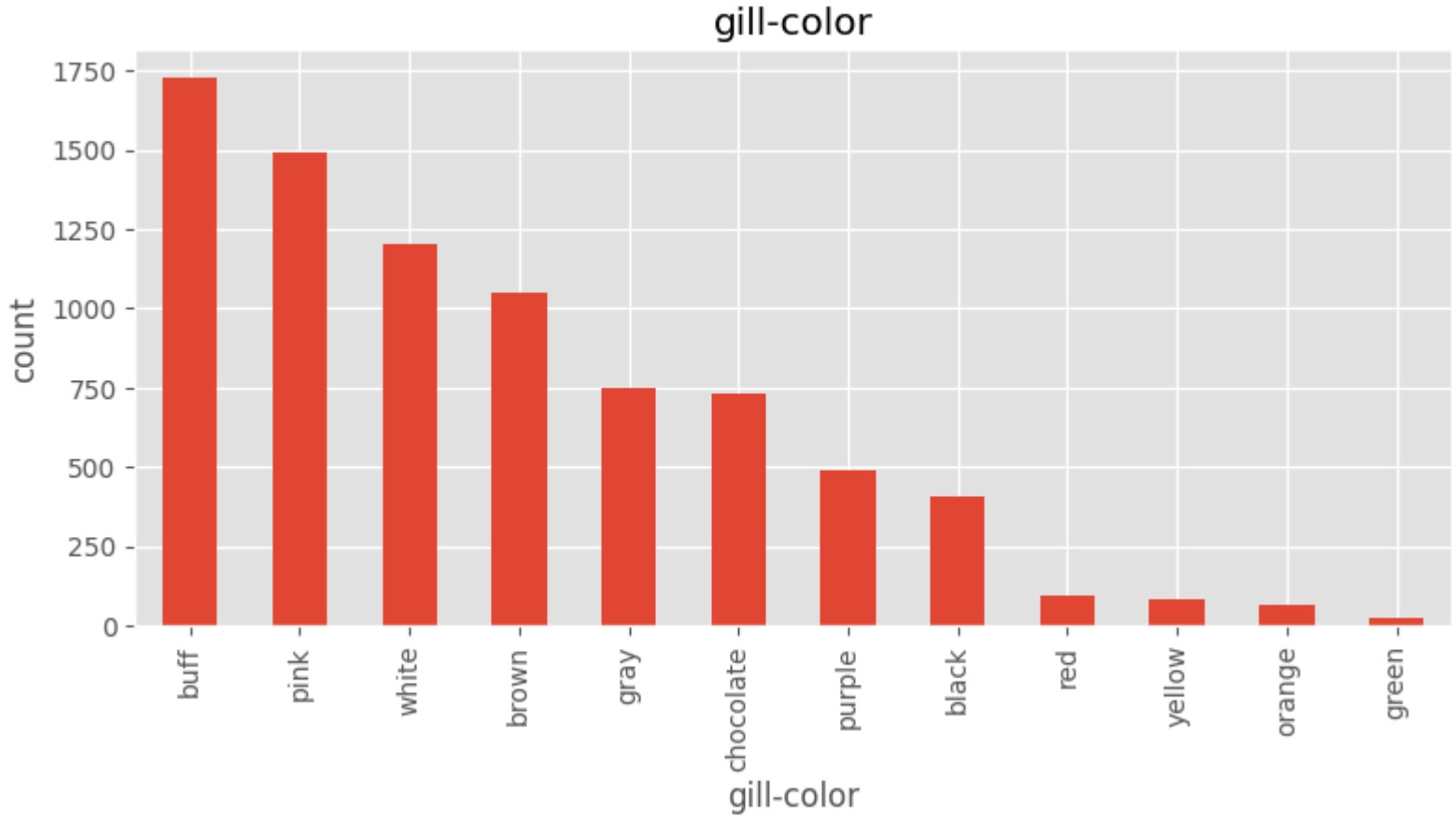

bruises
count
4000
3000
2000
1000
0
no
bruises
bruises

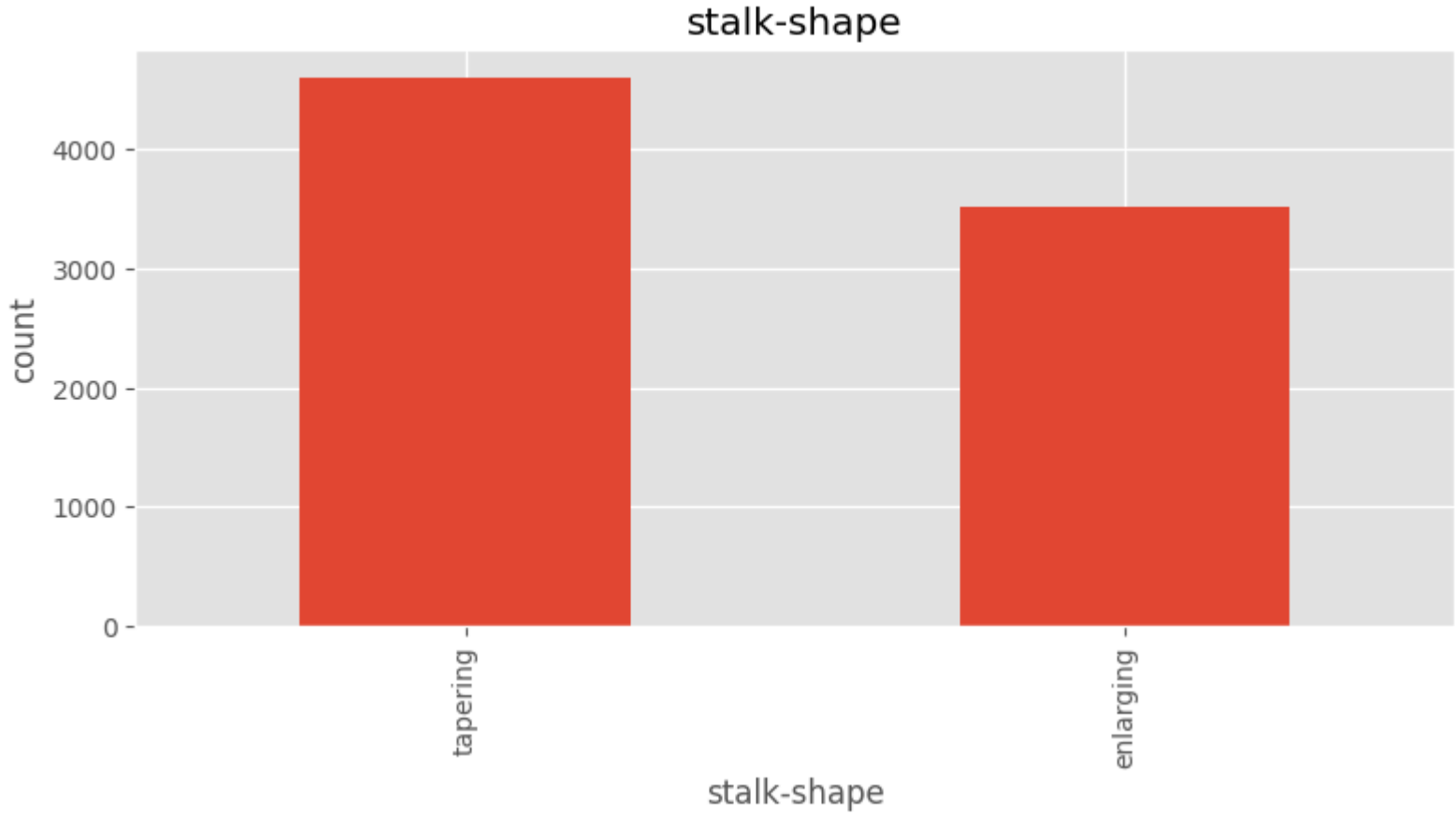

odor
count
3500
3000
2500
2000
1500
1000
500
0
none
foul
spicy
fishy
almond
anise
pungent
creosote
musty
odor

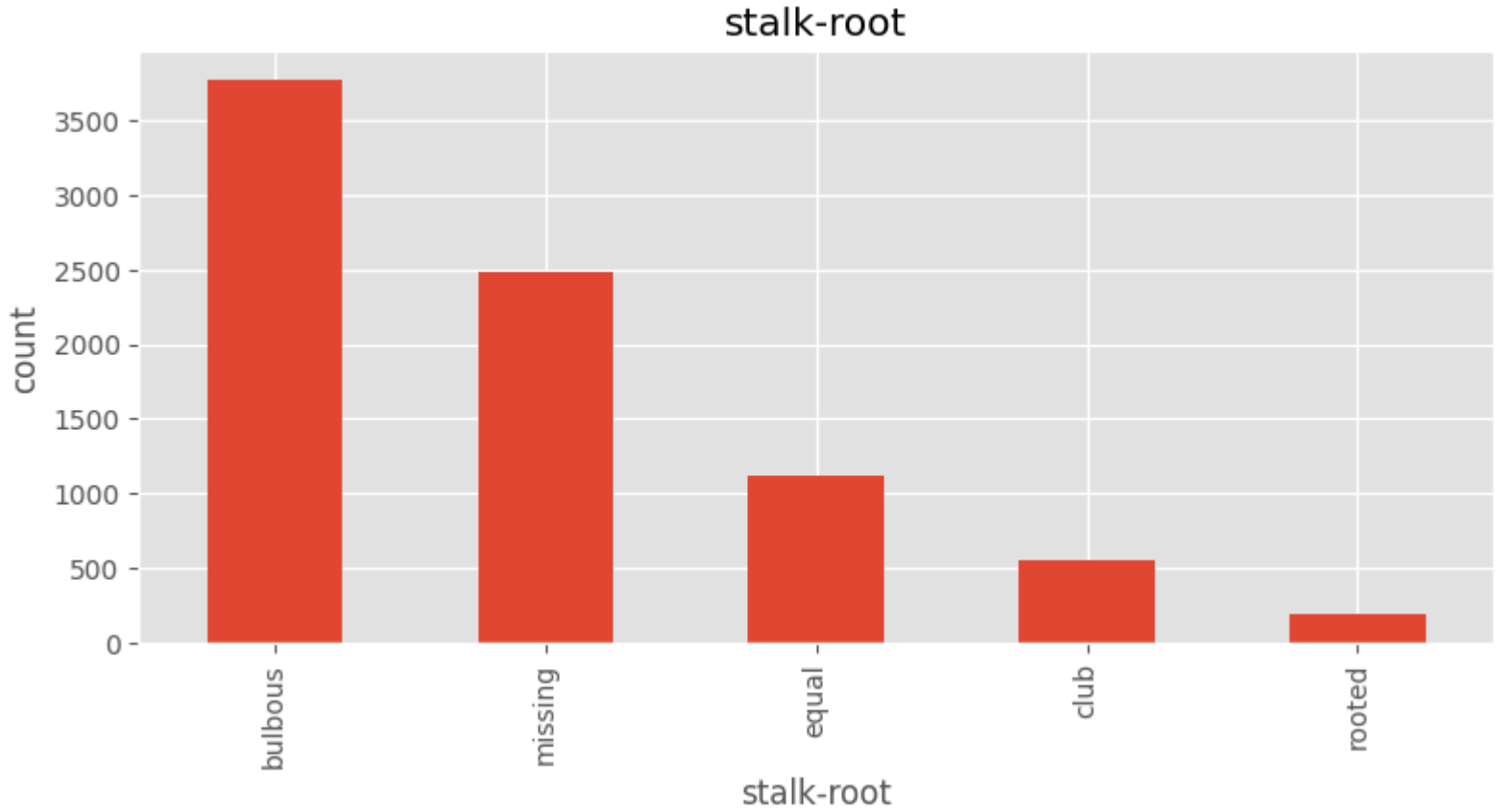

gill-attachment
8000
7000
6000
5000
4000
3000
2000
1000
0
count
free
attached
gill-attachment

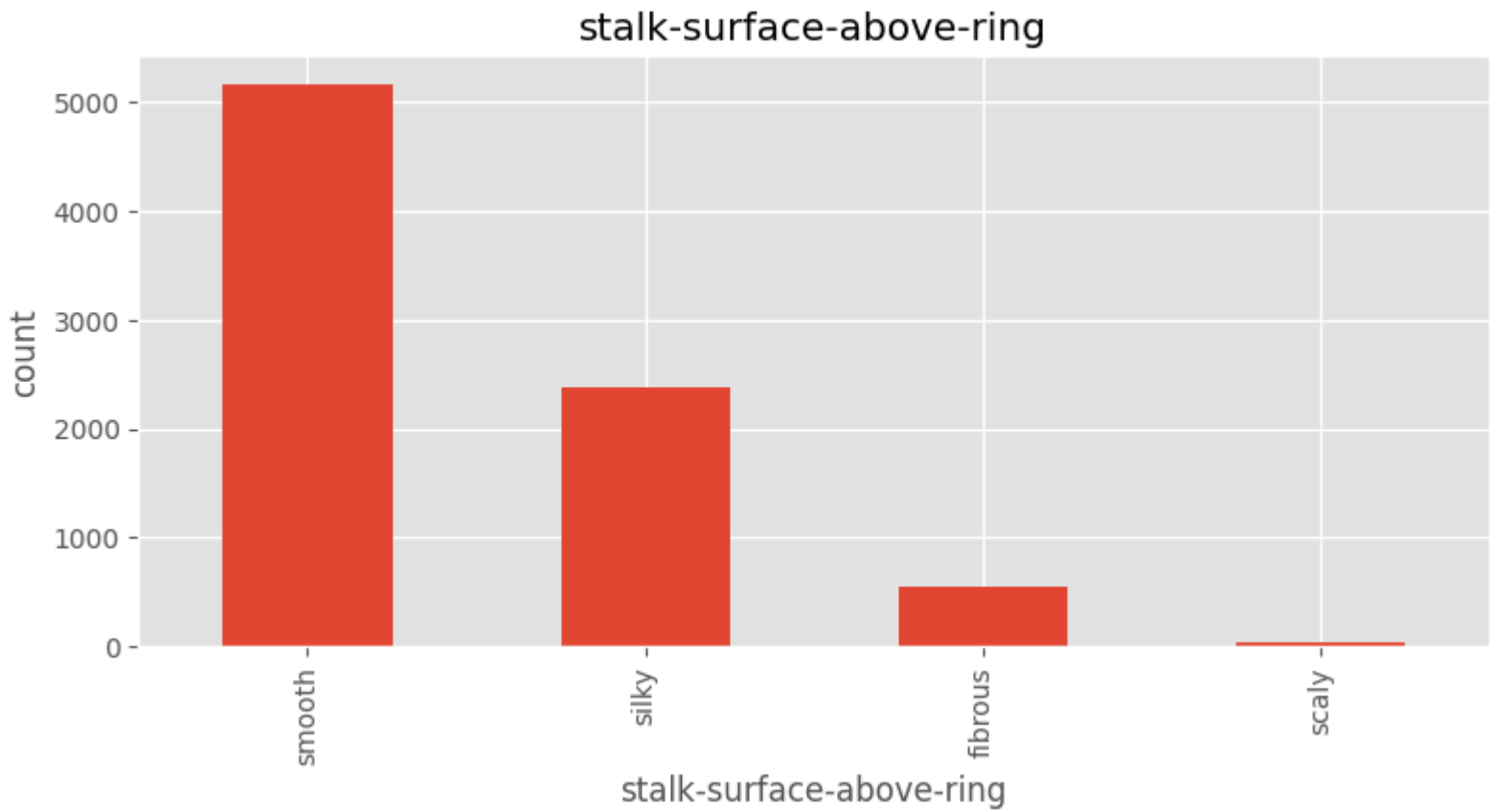

gill-spacing
7000
6000
5000
4000
3000
2000
1000
0
count
close
crowded
gill-spacing

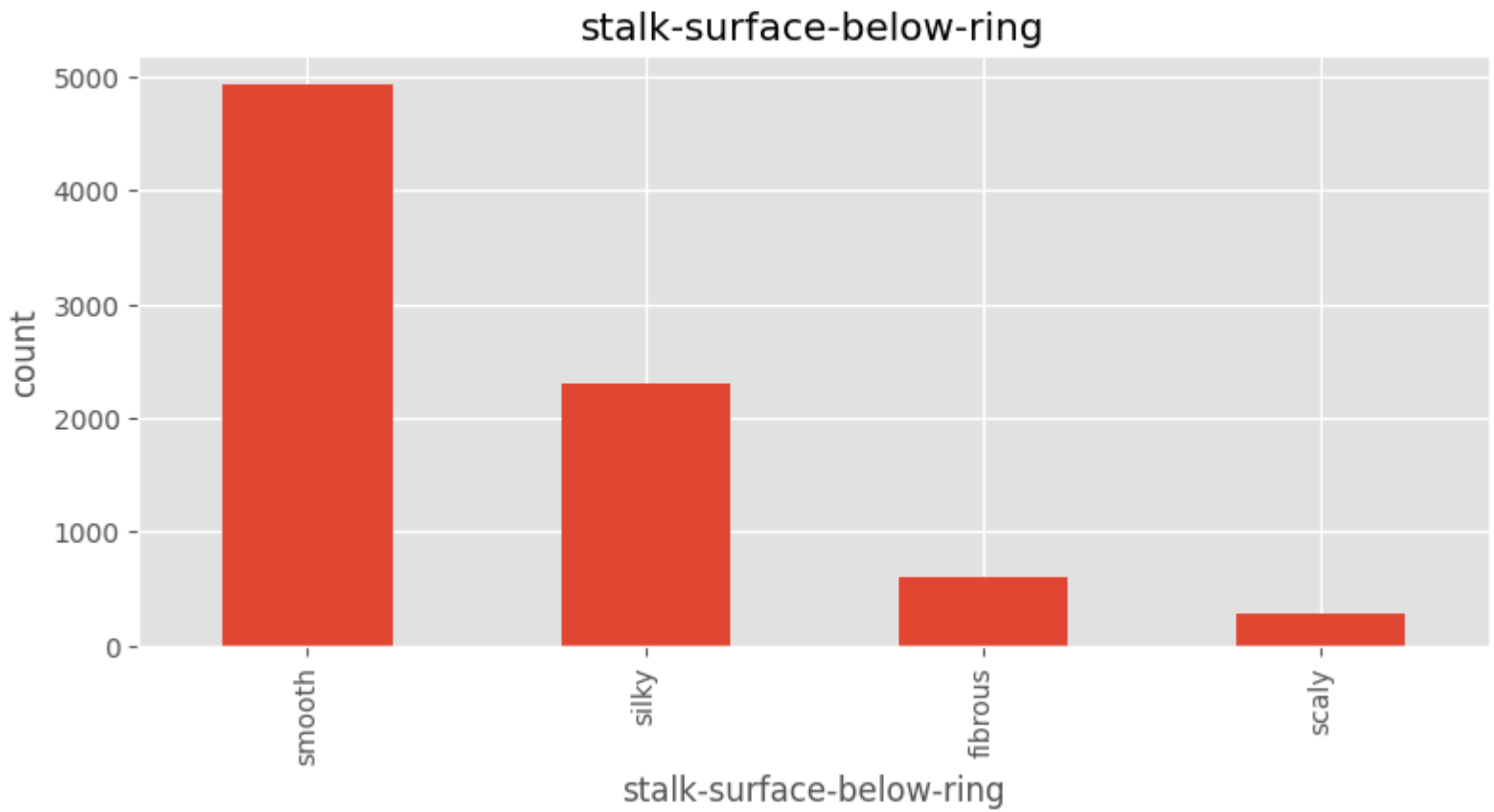

gill-size
count
5000
4000
3000
2000
1000
0
broad
narrow
gill-size

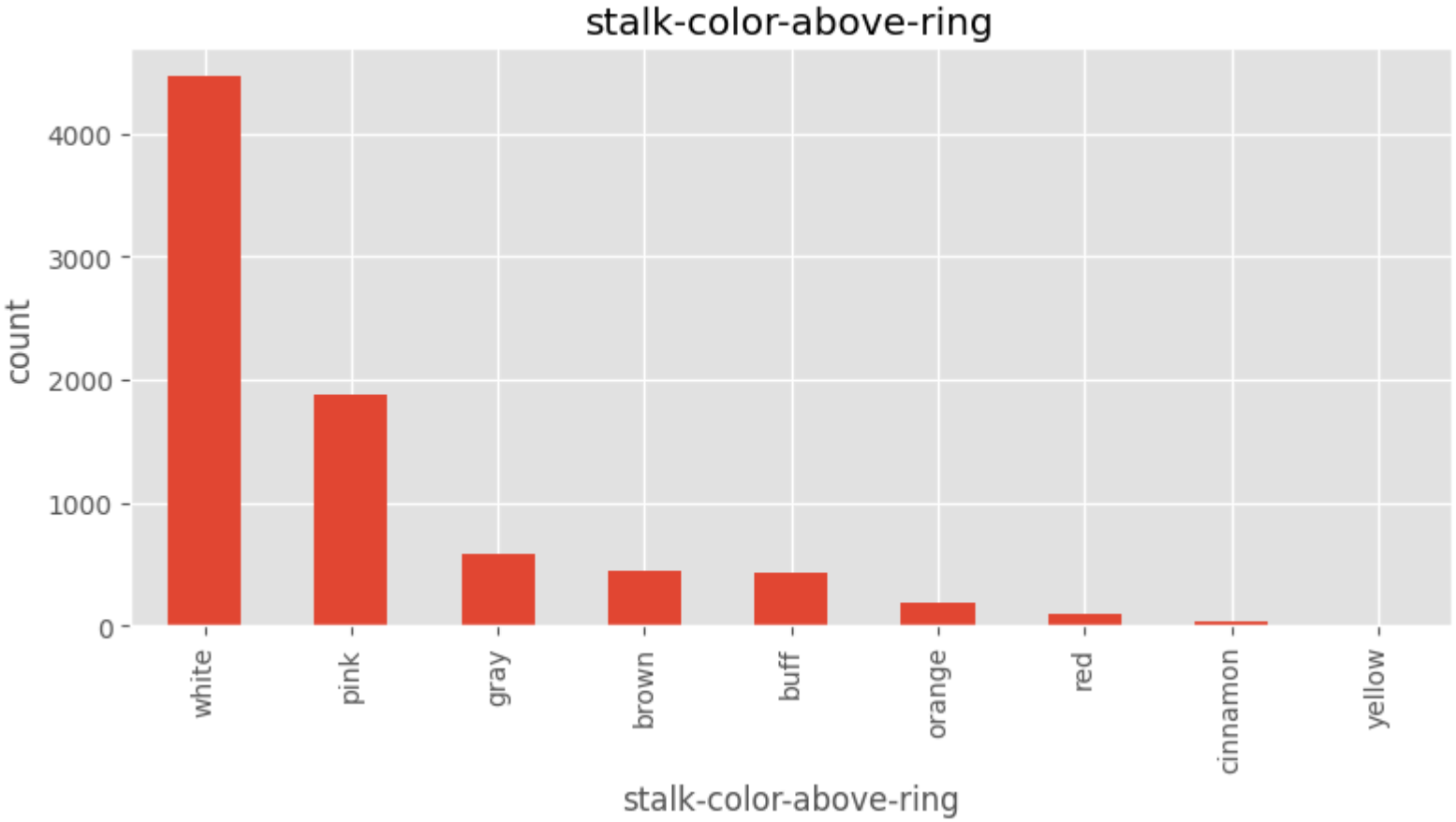

gill-color
count
1750
1500
1250
1000
750
500
250
0
buff
pink
white
brown
gray
chocolate
purple
black
red
yellow
orange
green
gill-color

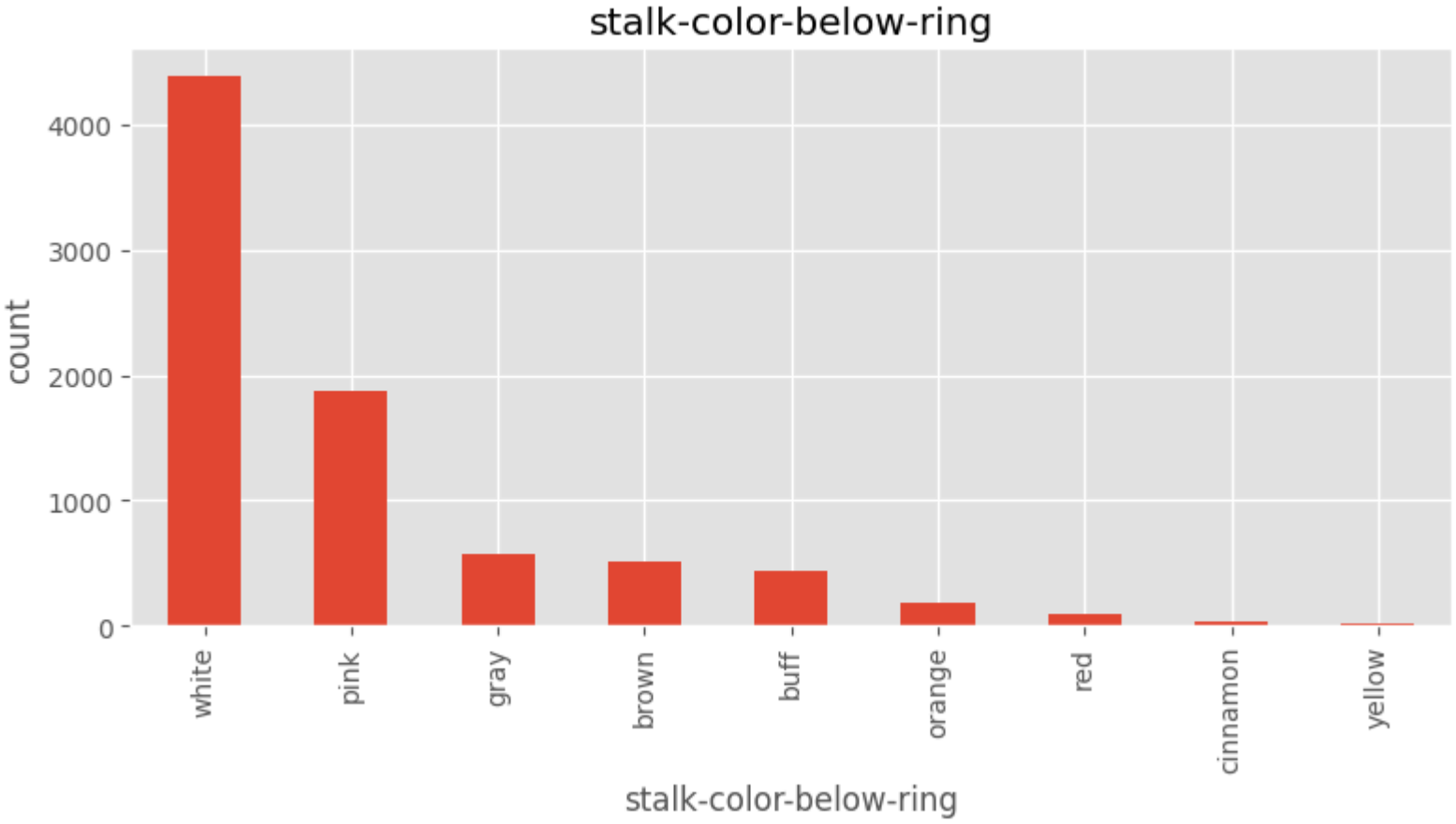

stalk-shape
count
4000
3000
2000
1000
0
tapering
enlarging
stalk-shape

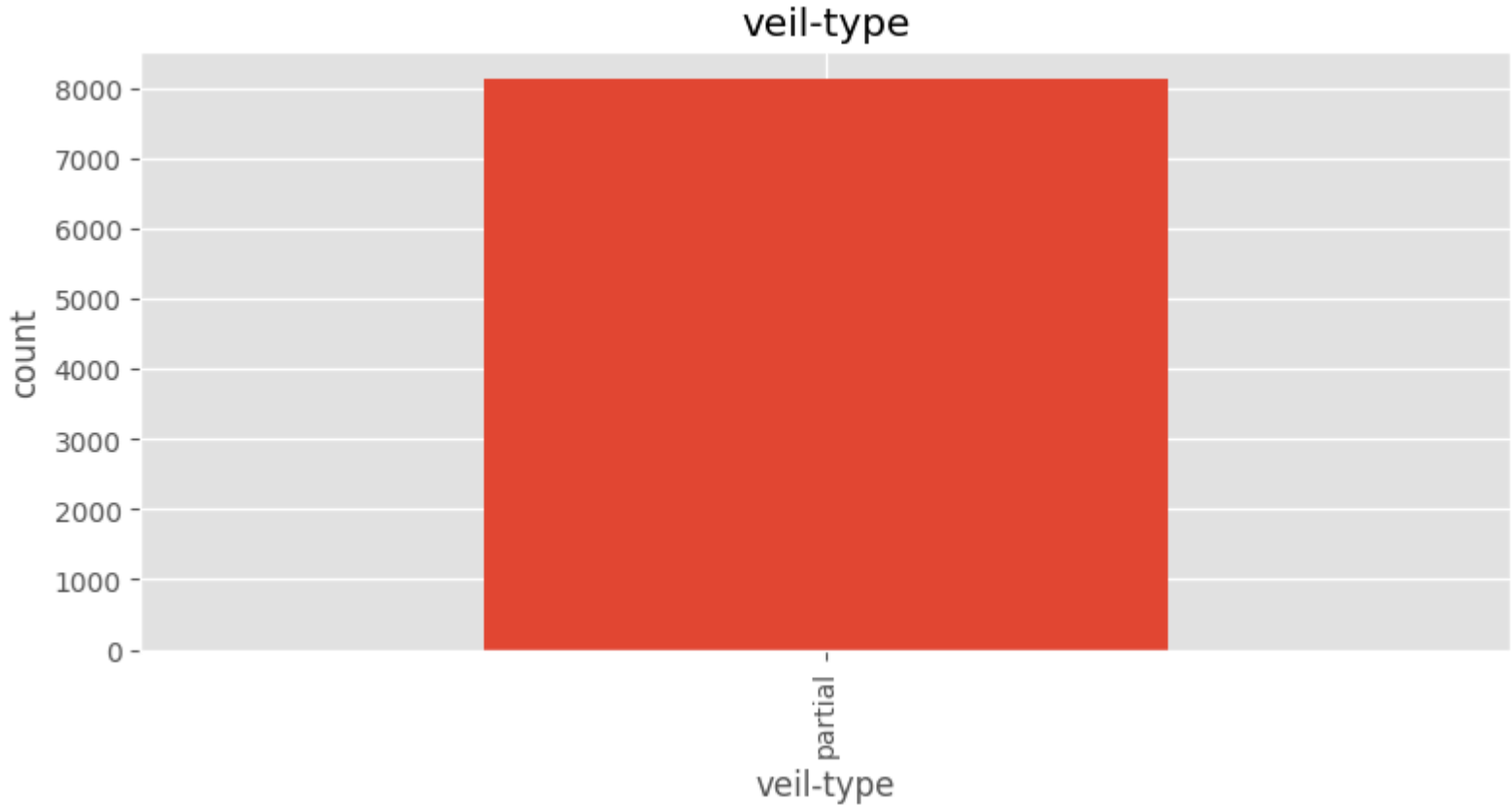

stalk-root
count
3500
3000
2500
2000
1500
1000
500
0
bulbous
missing
equal
club
rooted
stalk-root

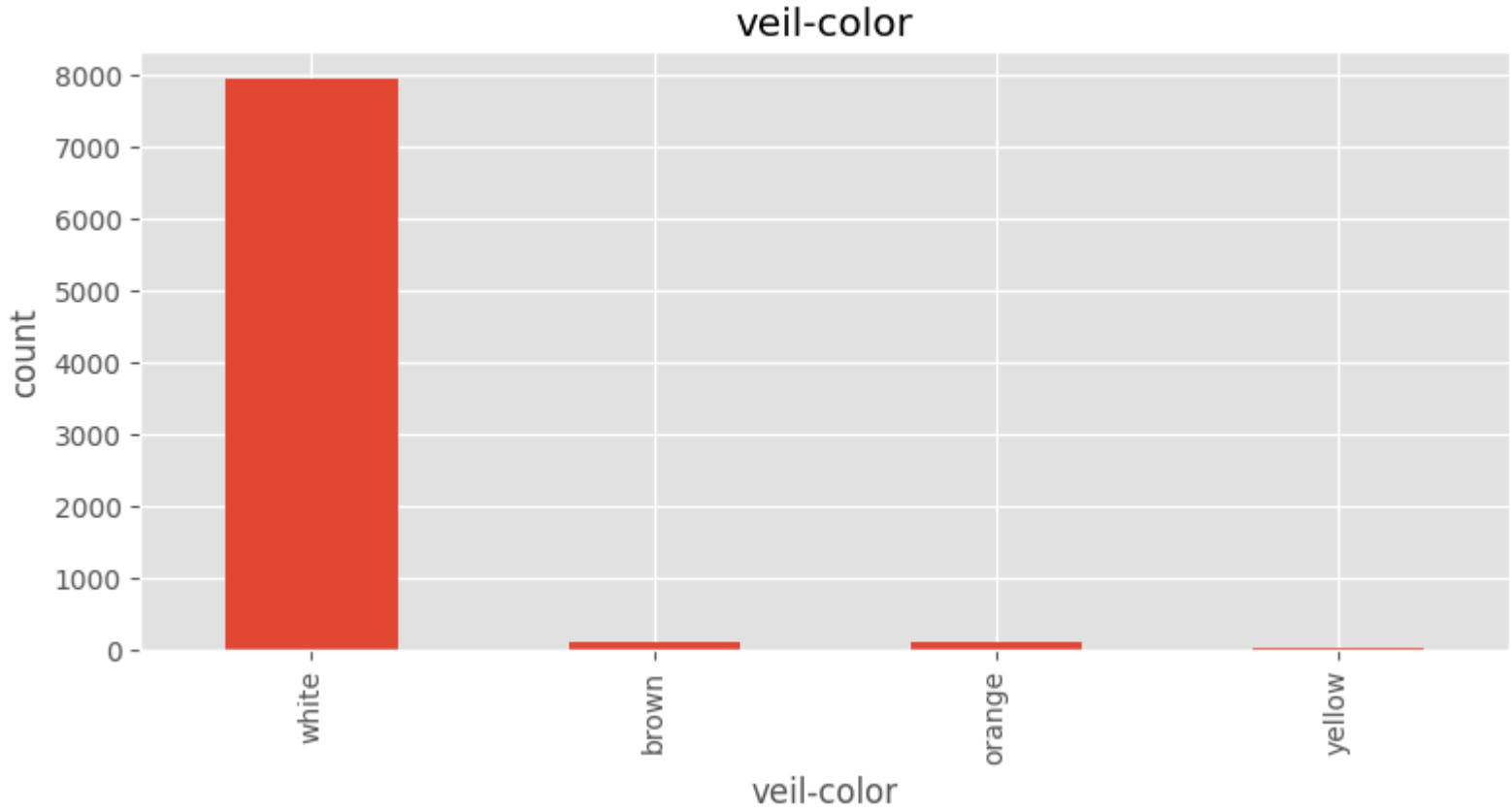

stalk-surface-above-ring
count
5000
4000
3000
2000
1000
0
smooth
silky
fibrous
scaly
stalk-surface-above-ring

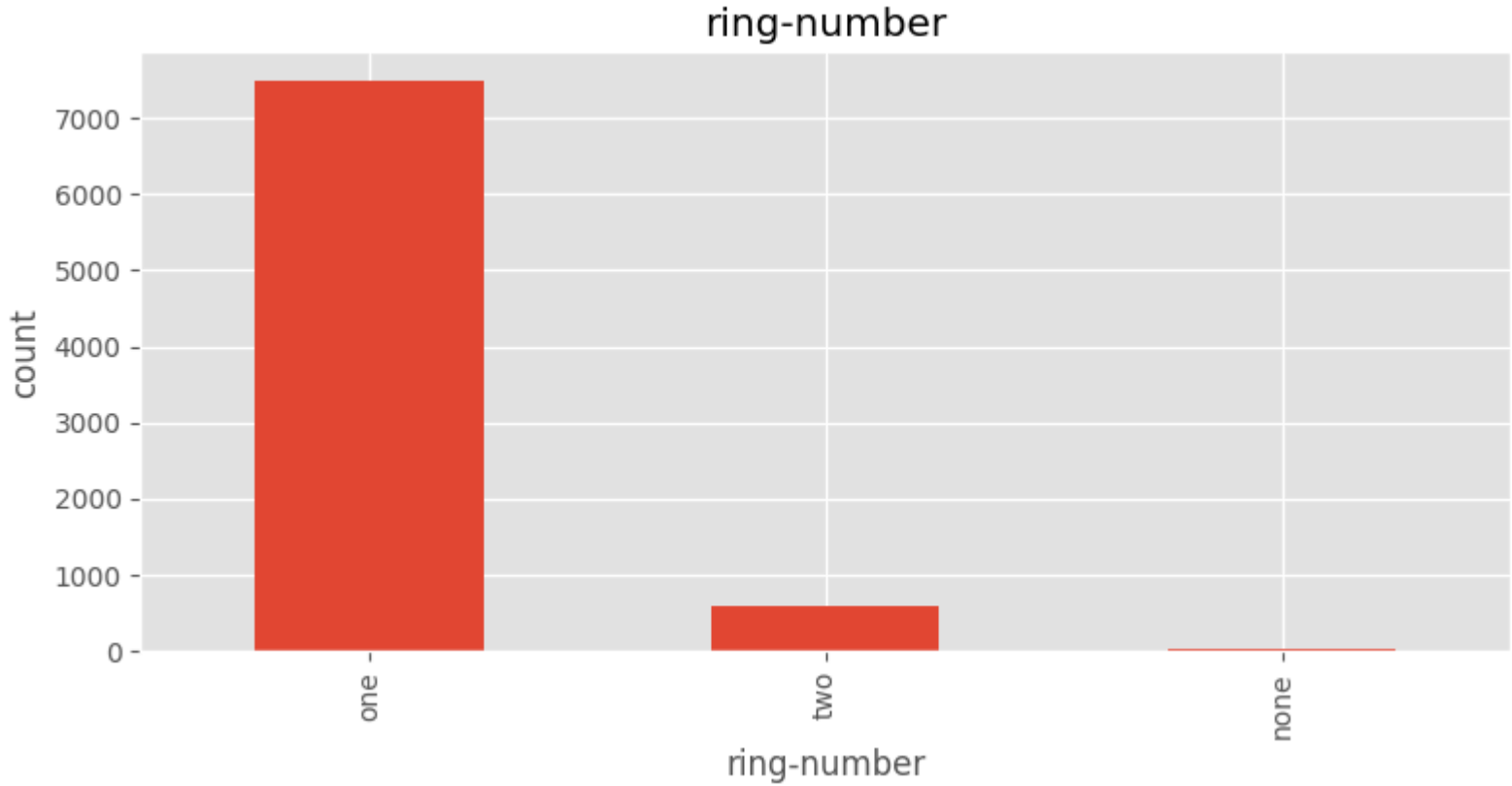

stalk-surface-below-ring
count
5000
4000
3000
2000
1000
0
smooth
silky
fibrous
scaly
stalk-surface-below-ring

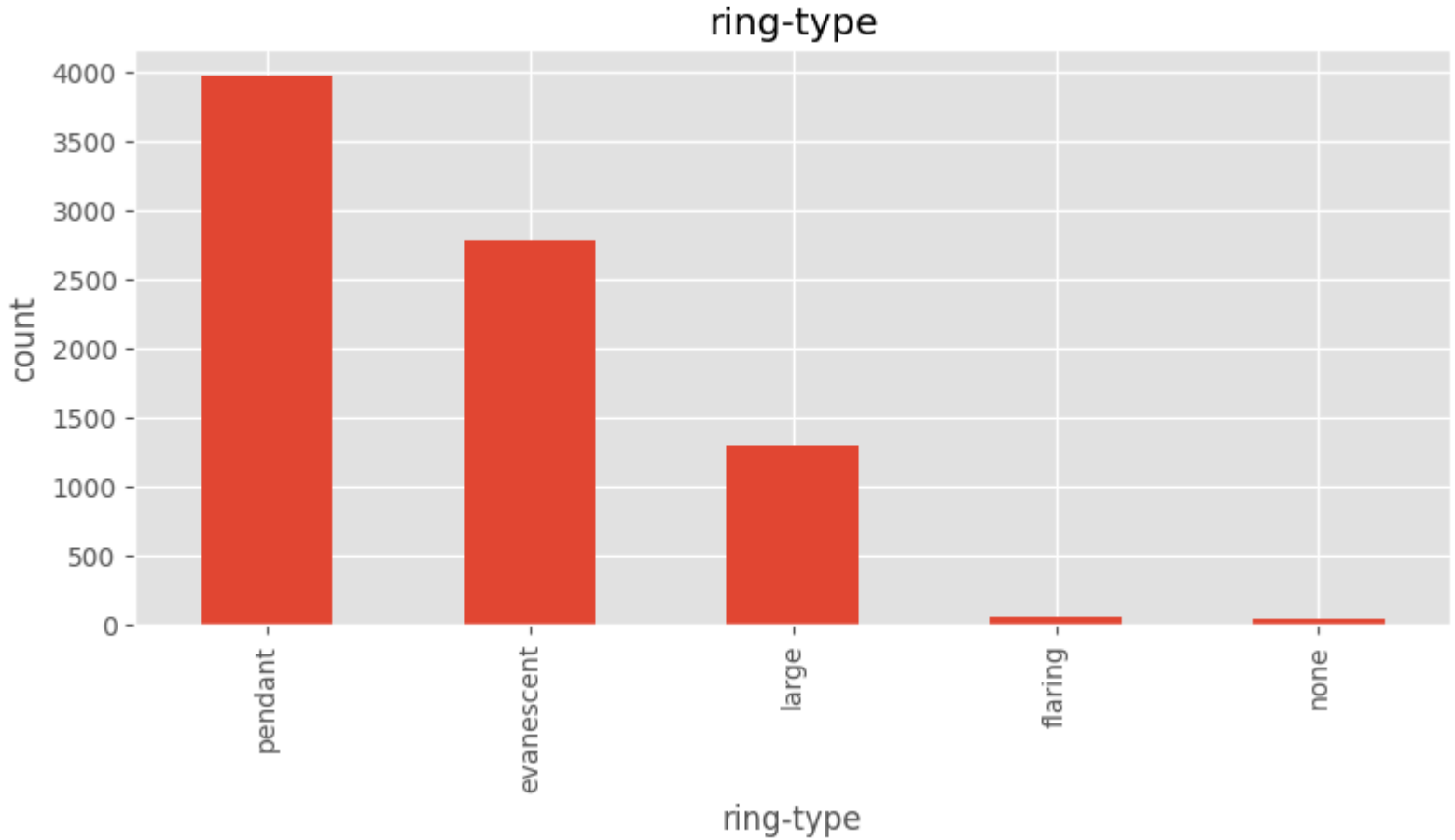

stalk-color-above-ring
count
4000
3000
2000
1000
0
white
pink
gray
brown
buff
orange
red
cinnamon
yellow
stalk-color-above-ring

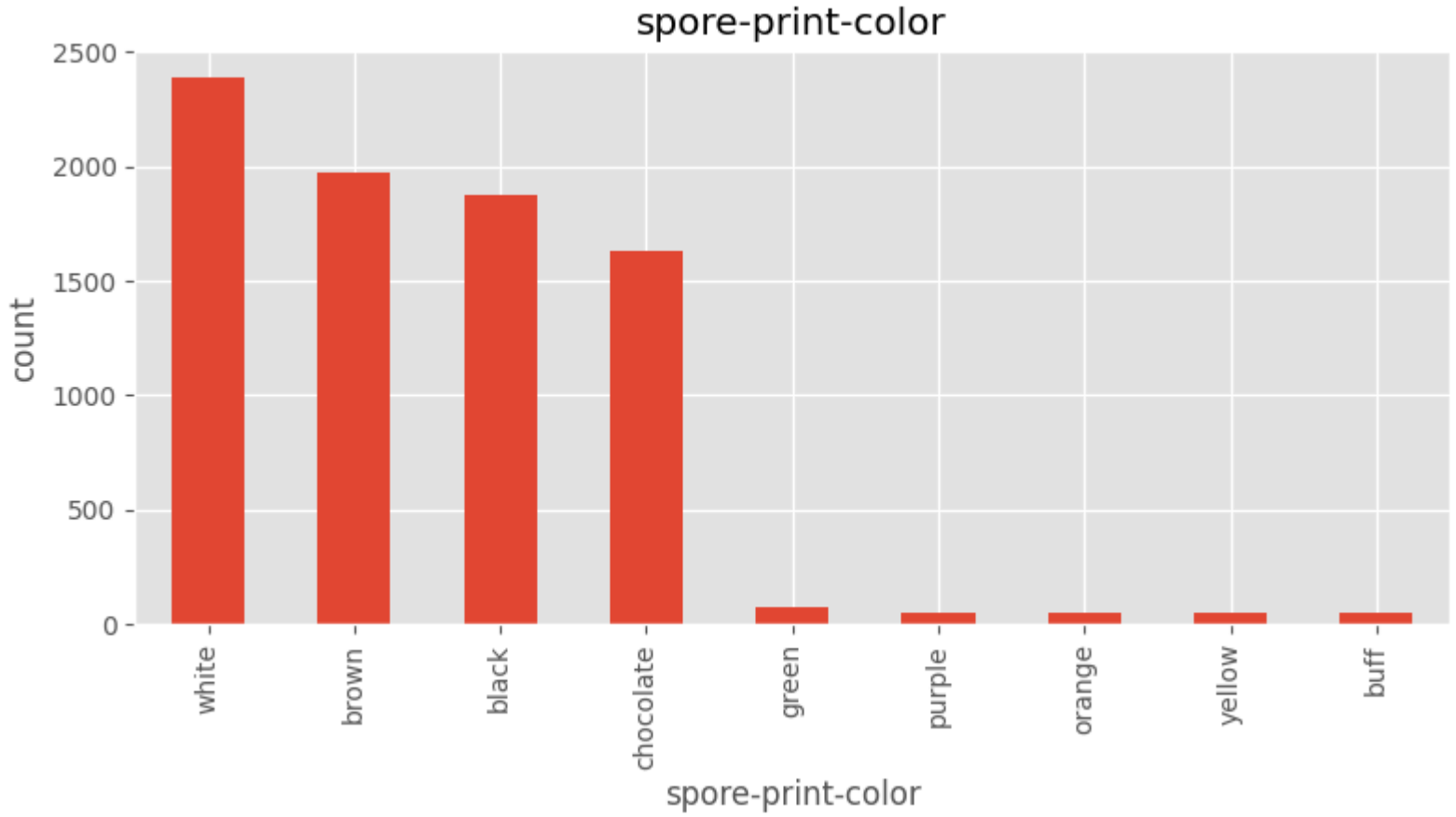

stalk-color-below-ring
count
4000
3000
2000
1000
0
white
pink
gray
brown
buff
orange
red
cinnamon
yellow
stalk-color-below-ring

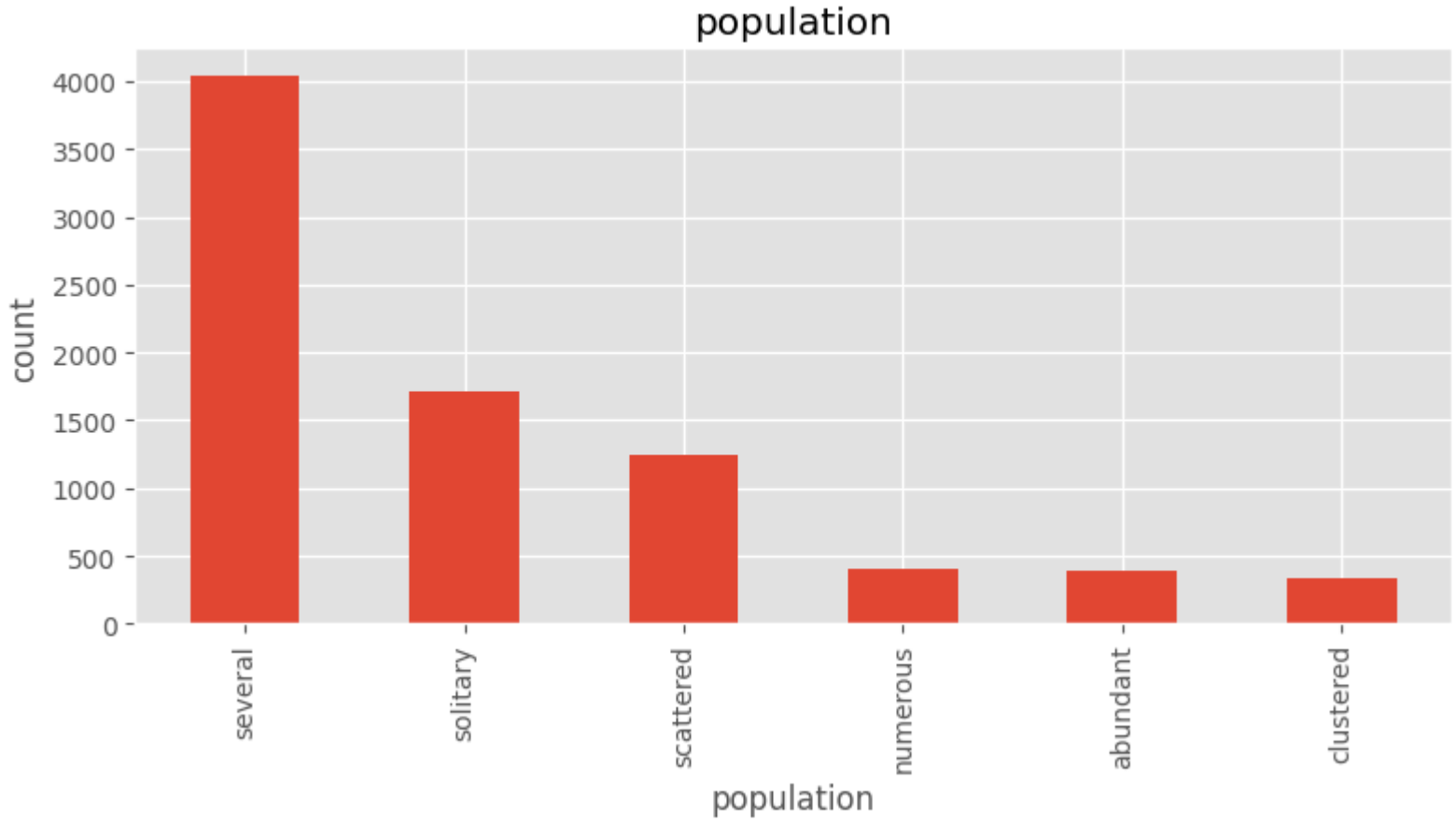

veil-type
8000
7000
6000
5000
4000
3000
2000
1000
0
count
partial
veil-type

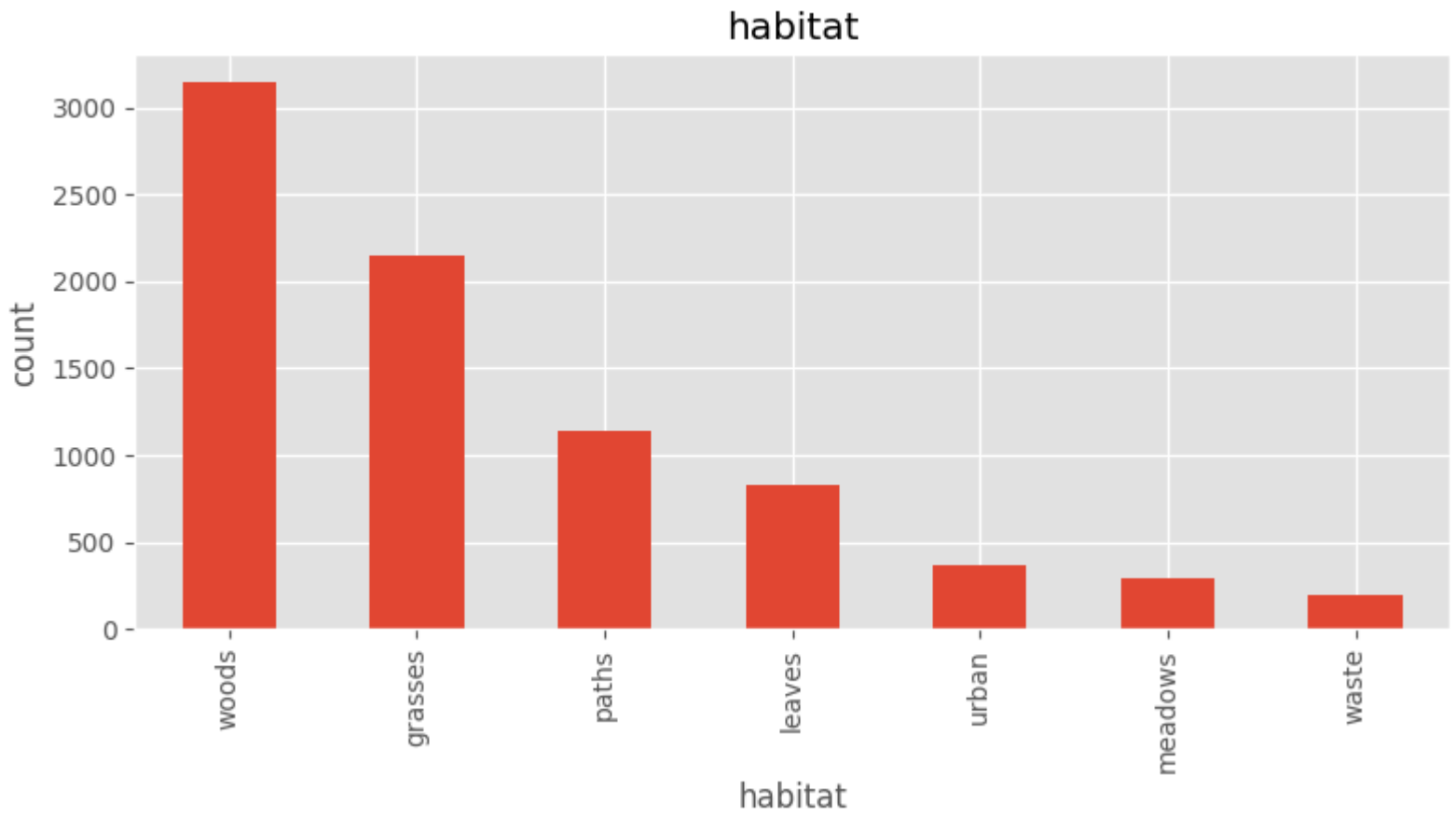

veil-color
8000
7000
6000
5000
4000
3000
2000
1000
0
count
white
brown
orange
yellow
veil-color


ring-number
count
7000
6000
5000
4000
3000
2000
1000
0
one
two
none
ring-number


ring-type
count
4000
3500
3000
2500
2000
1500
1000
500
0
pendant
evanescent
large
flaring
none
ring-type

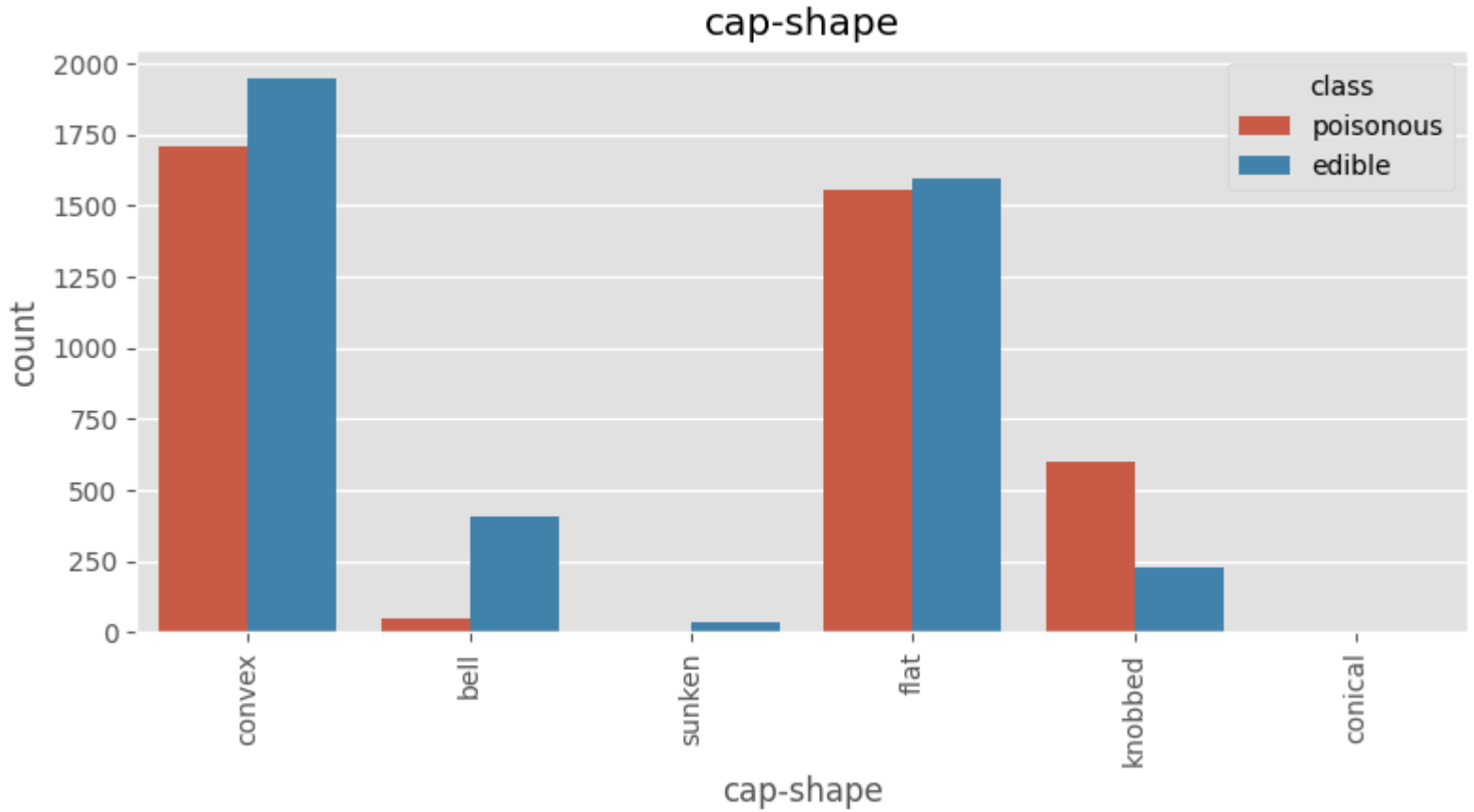

spore-print-color
2500
2000
1500
1000
500
0
count
white
brown
black
chocolate
green
purple
orange
yellow
buff
spore-print-color

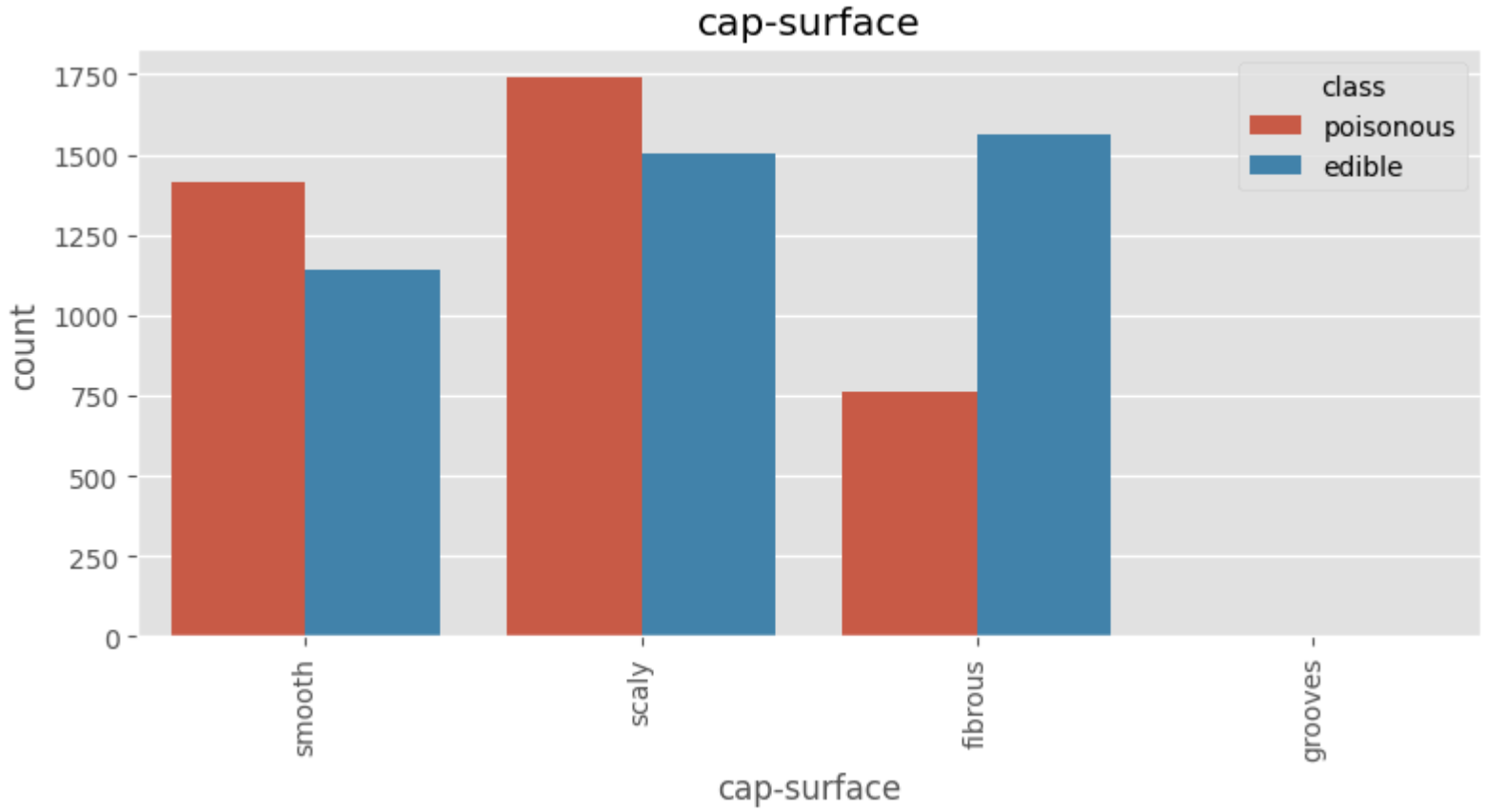

population
count
4000
3500
3000
2500
2000
1500
1000
500
0
several
solitary
scattered
numerous
abundant
clustered
population

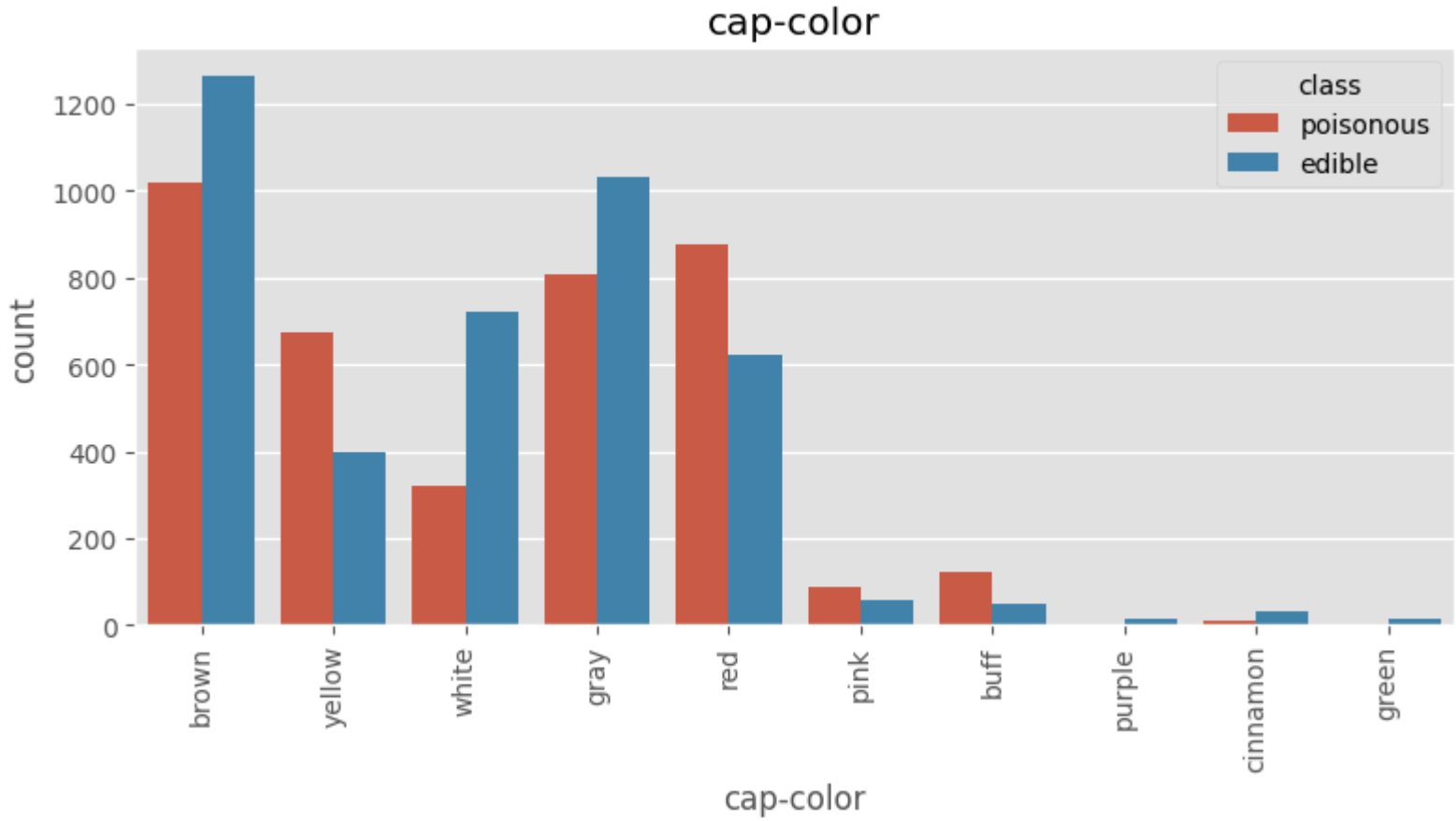


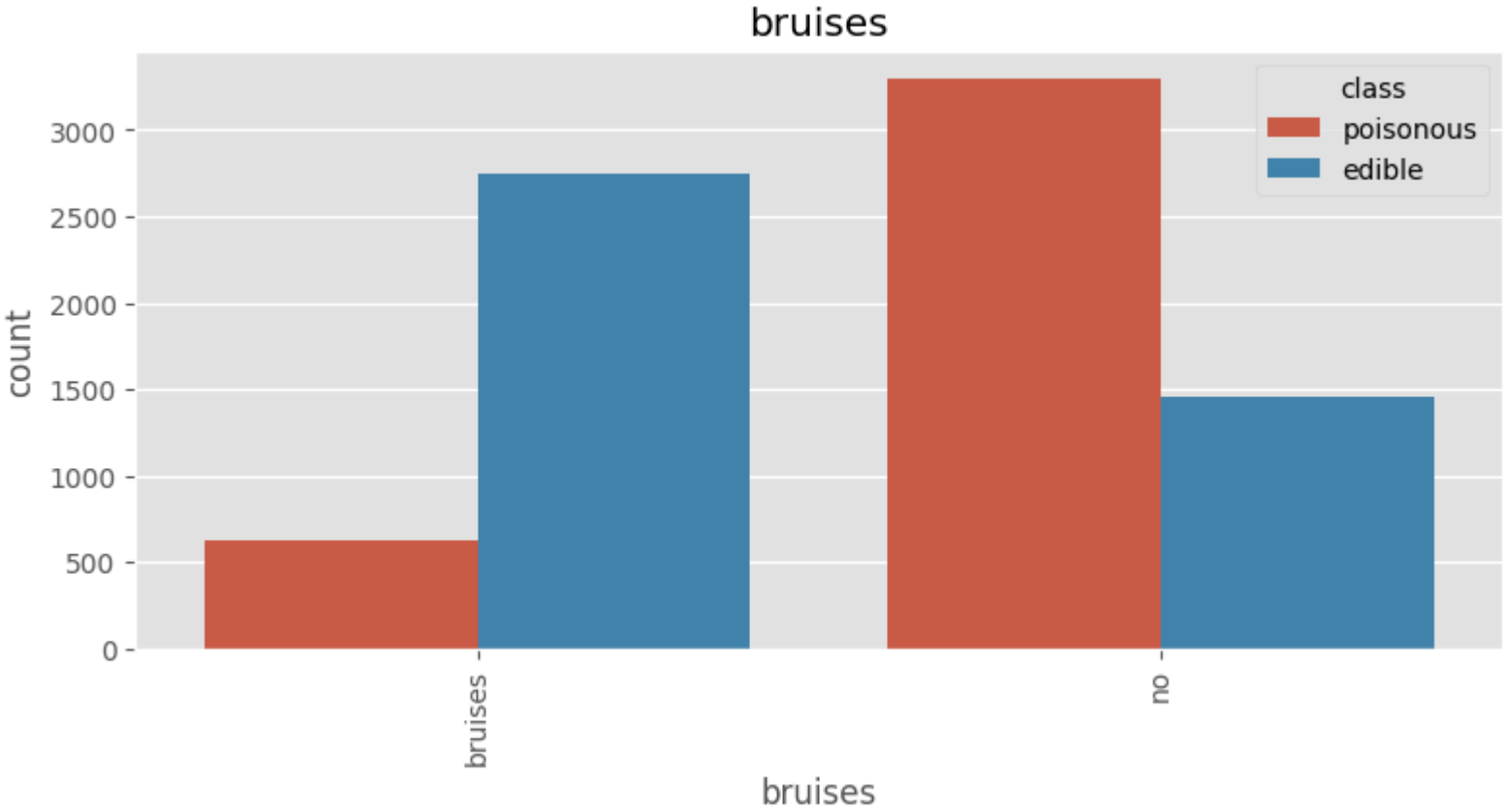
In [ ]:
```
# Bivariate analysis of features with the target variable
for column in mushrooms:
    if column == "class":
        continue
    plt.figure(figsize=(20,4))
    plt.subplot(121)
    sns.countplot(mushrooms, x=column, hue="class")
    plt.title(column)
    plt.xticks(rotation=90)
```

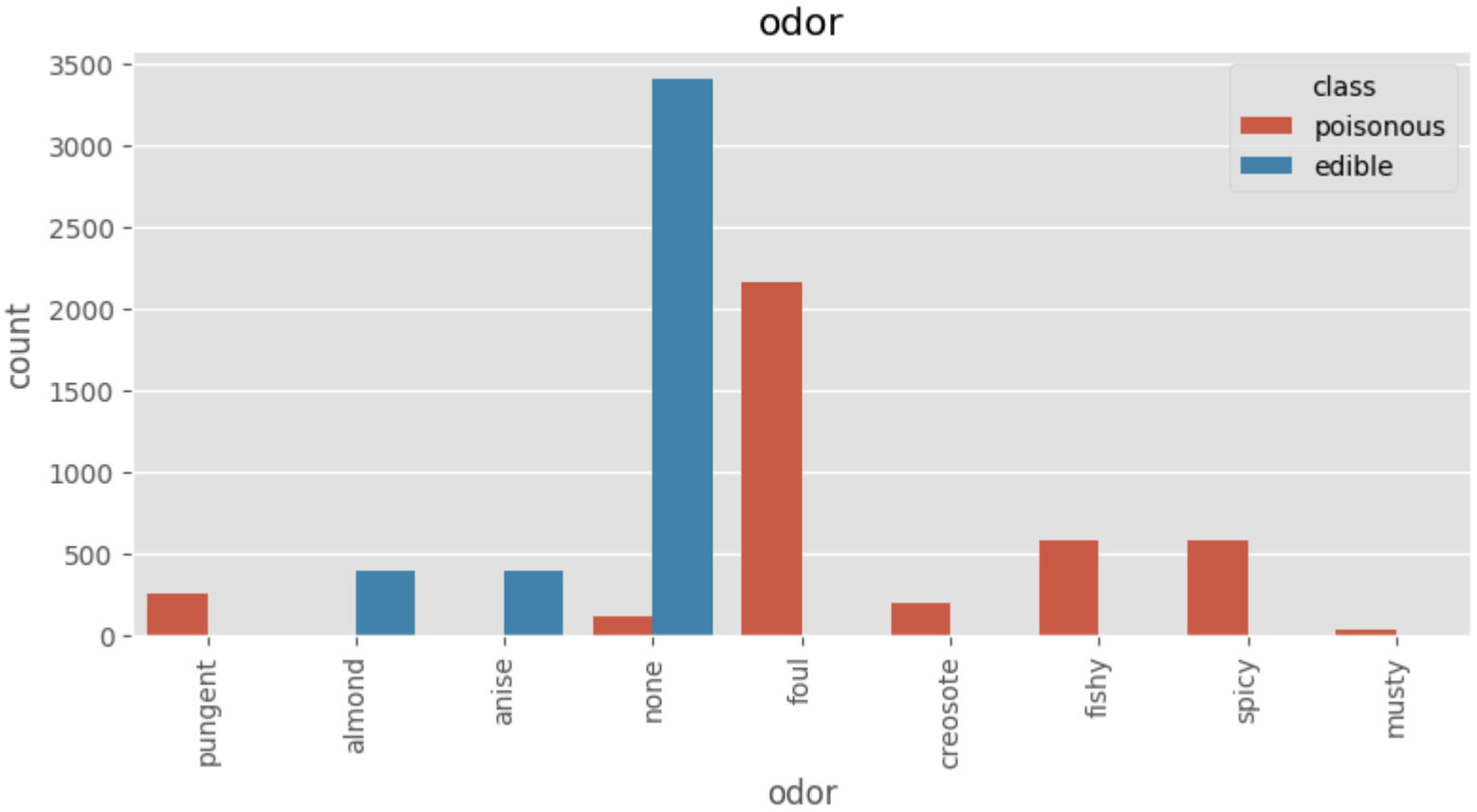
```
<ipython-input-115-44140a7c0a7a>:6: RuntimeWarning: More than 20 figures have been opened. Figures created through the pyplot i
nterface (`matplotlib.pyplot.figure`) are retained until explicitly closed and may consume too much memory. (To control this wa
rning, see the rcParam `figure.max_open_warning`). Consider using `matplotlib.pyplot.close()`.
  plt.figure(figsize=(20,4))
```

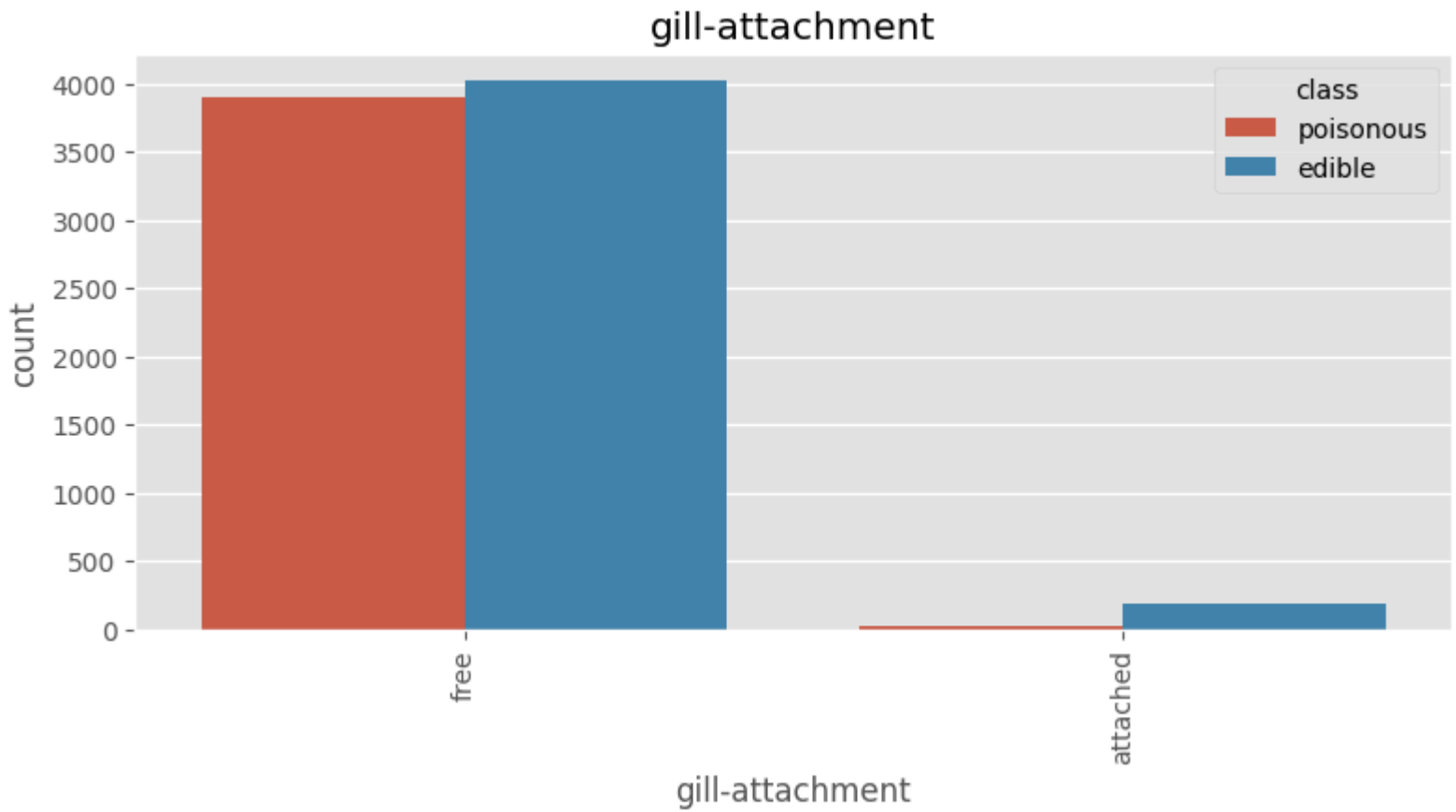


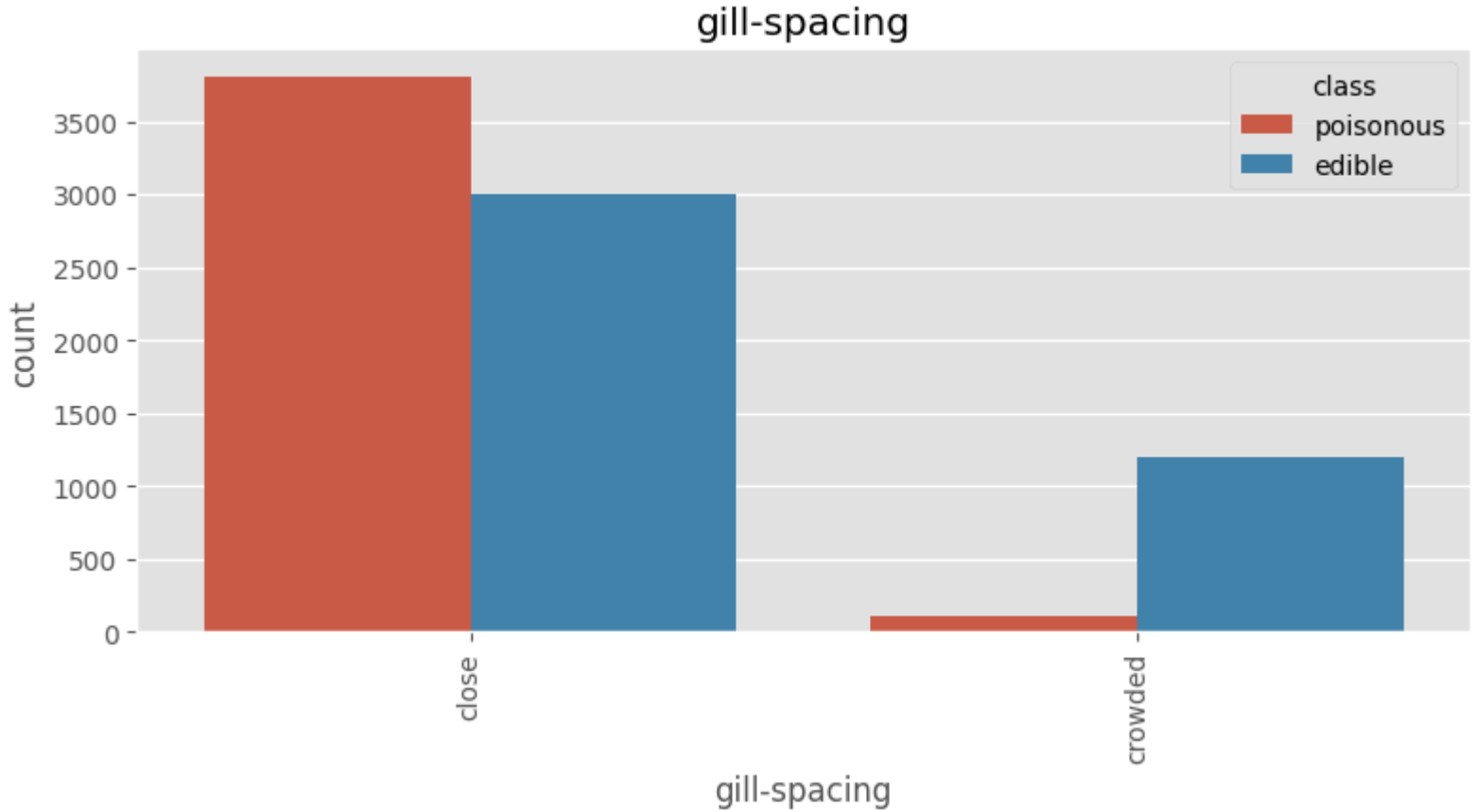

cap-surface
class
poisonous
edible
count
1750
1500
1250
1000
750
500
250
0
smooth
scaly
fibrous
grooves
cap-surface

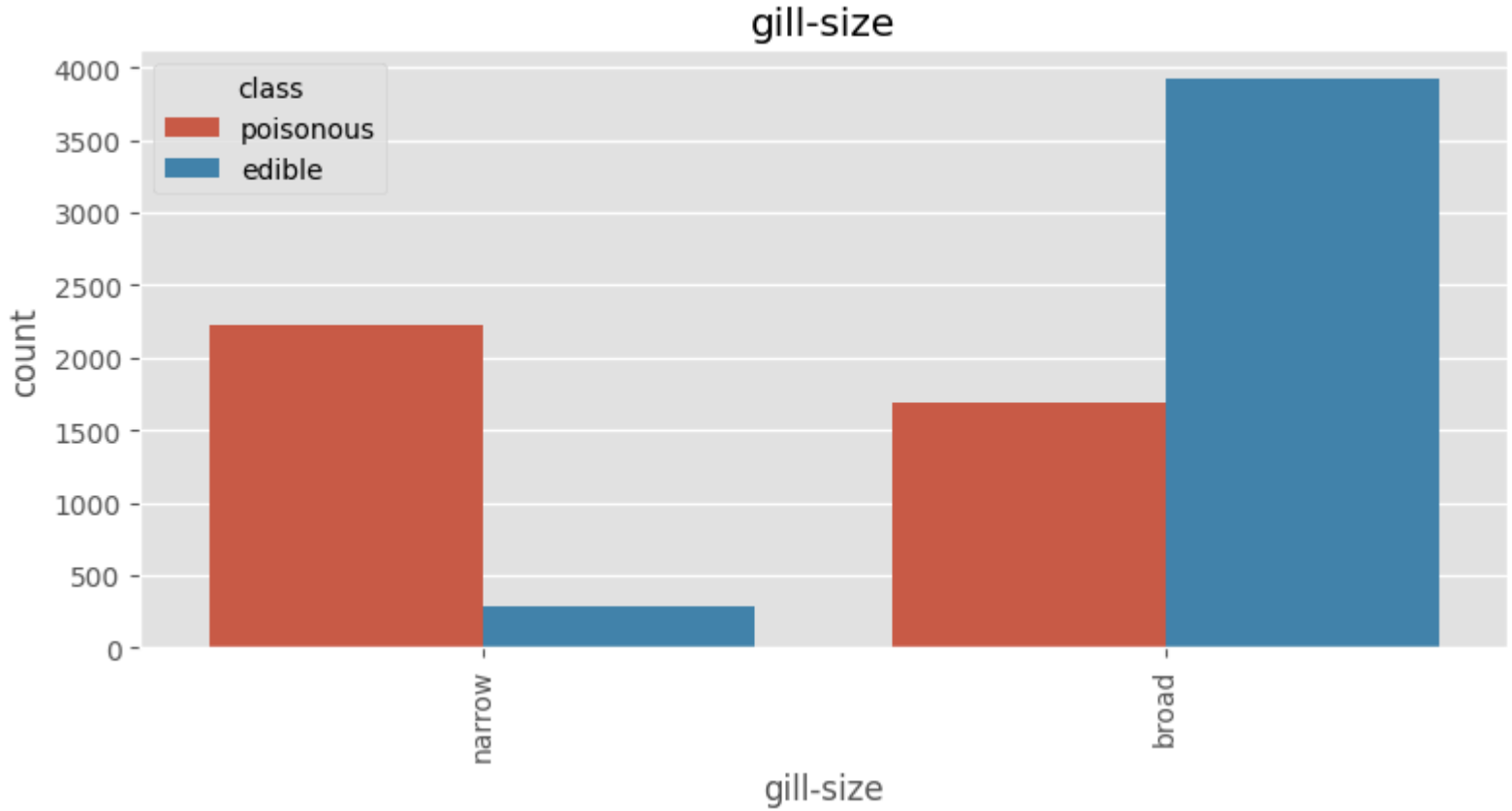

cap-color
class
poisonous
edible
count
1200
1000
800
600
400
200
0
brown
yellow
white
gray
red
pink
buff
purple
cinnamon
green
cap-color

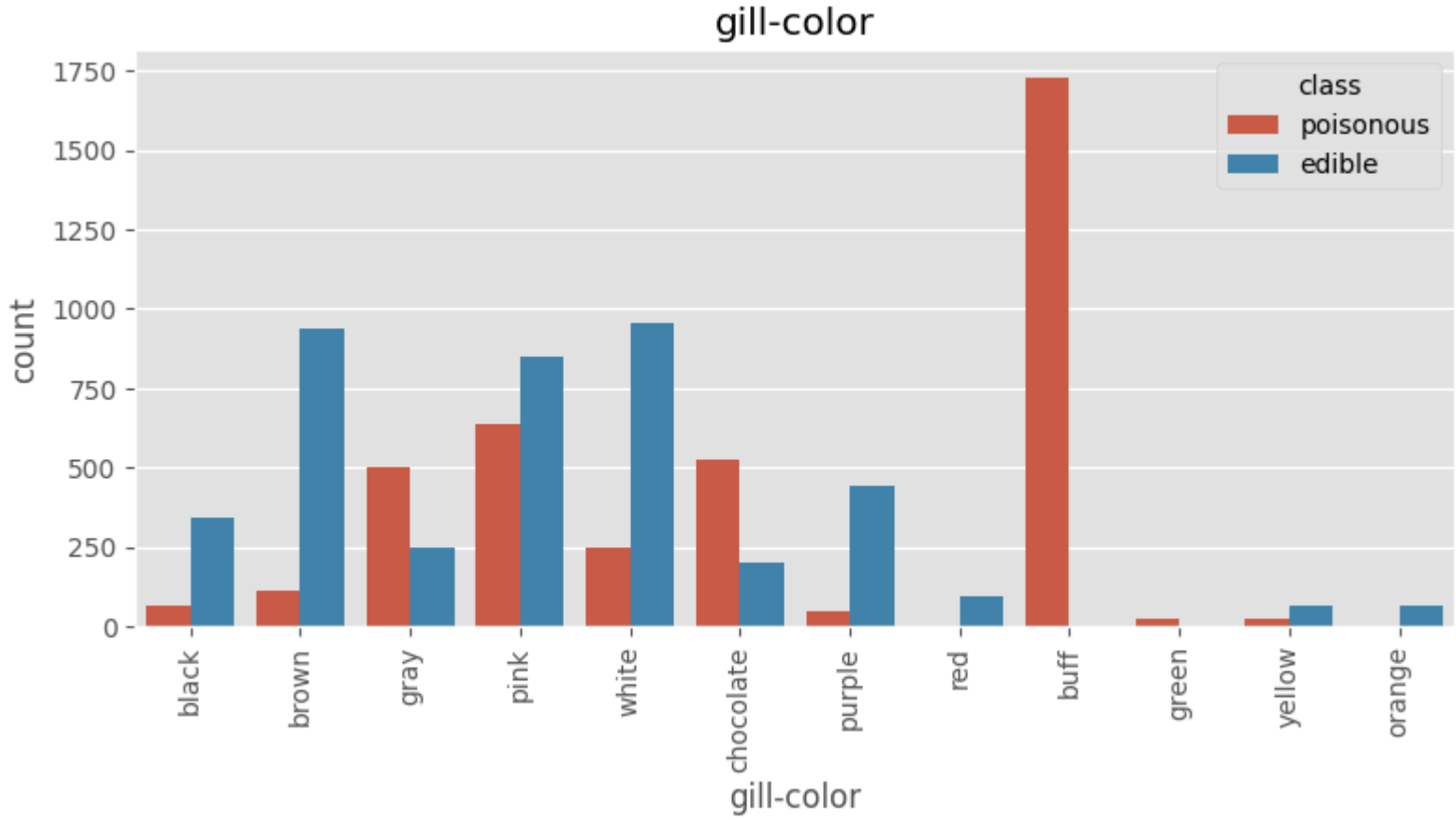

bruises
class
poisonous
edible
count
3000
2500
2000
1500
1000
500
0
bruises
no
bruises

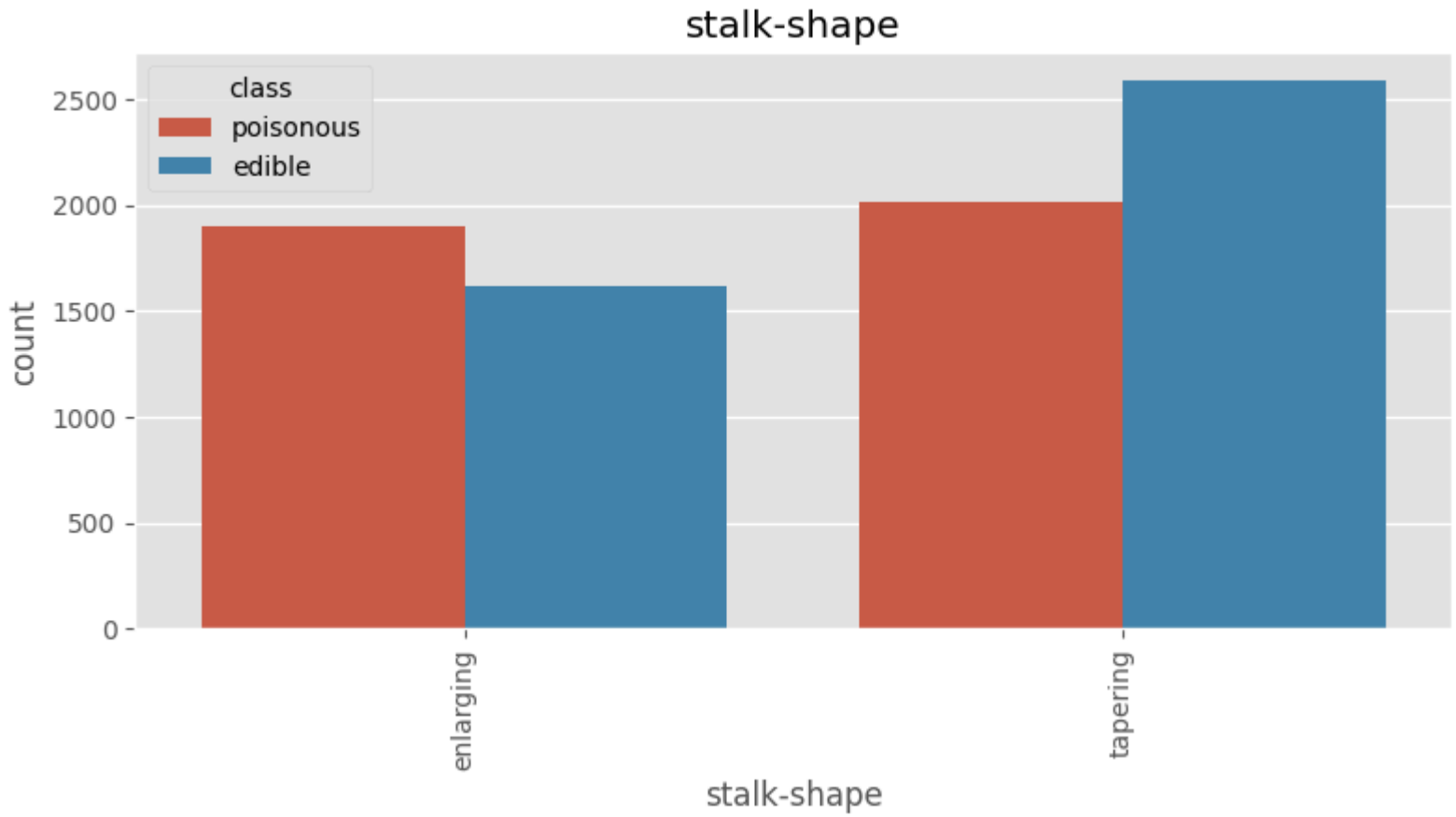

odor
class
poisonous
edible
count
3500
3000
2500
2000
1500
1000
500
0
pungent
almond
anise
none
foul
creosote
fishy
spicy
musty
odor

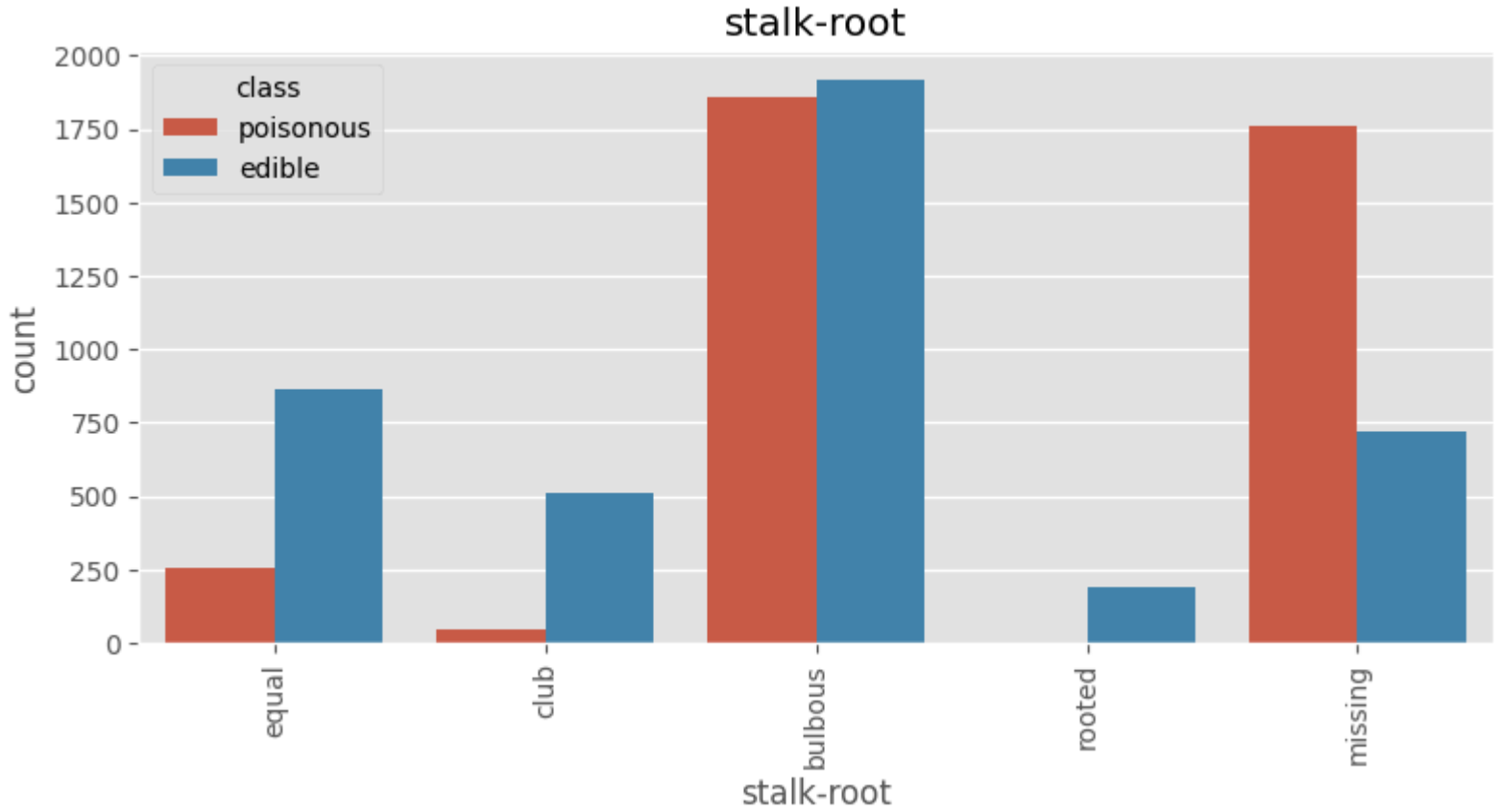

gill-attachment
class
poisonous
edible
count
4000
3500
3000
2500
2000
1500
1000
500
0
free
attached
gill-attachment

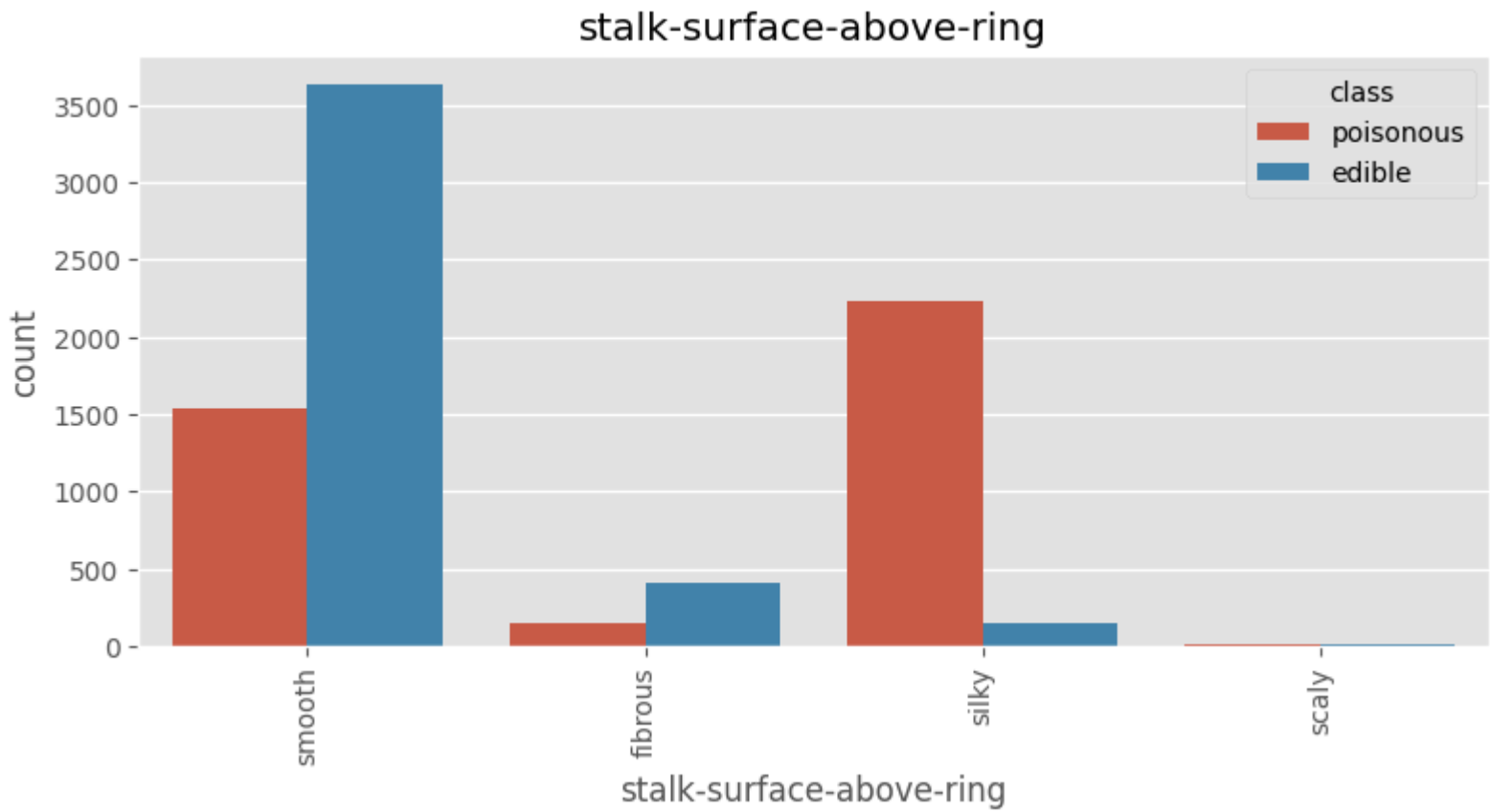

gill-spacing
class
poisonous
edible
count
3500
3000
2500
2000
1500
1000
500
0
close
crowded
gill-spacing

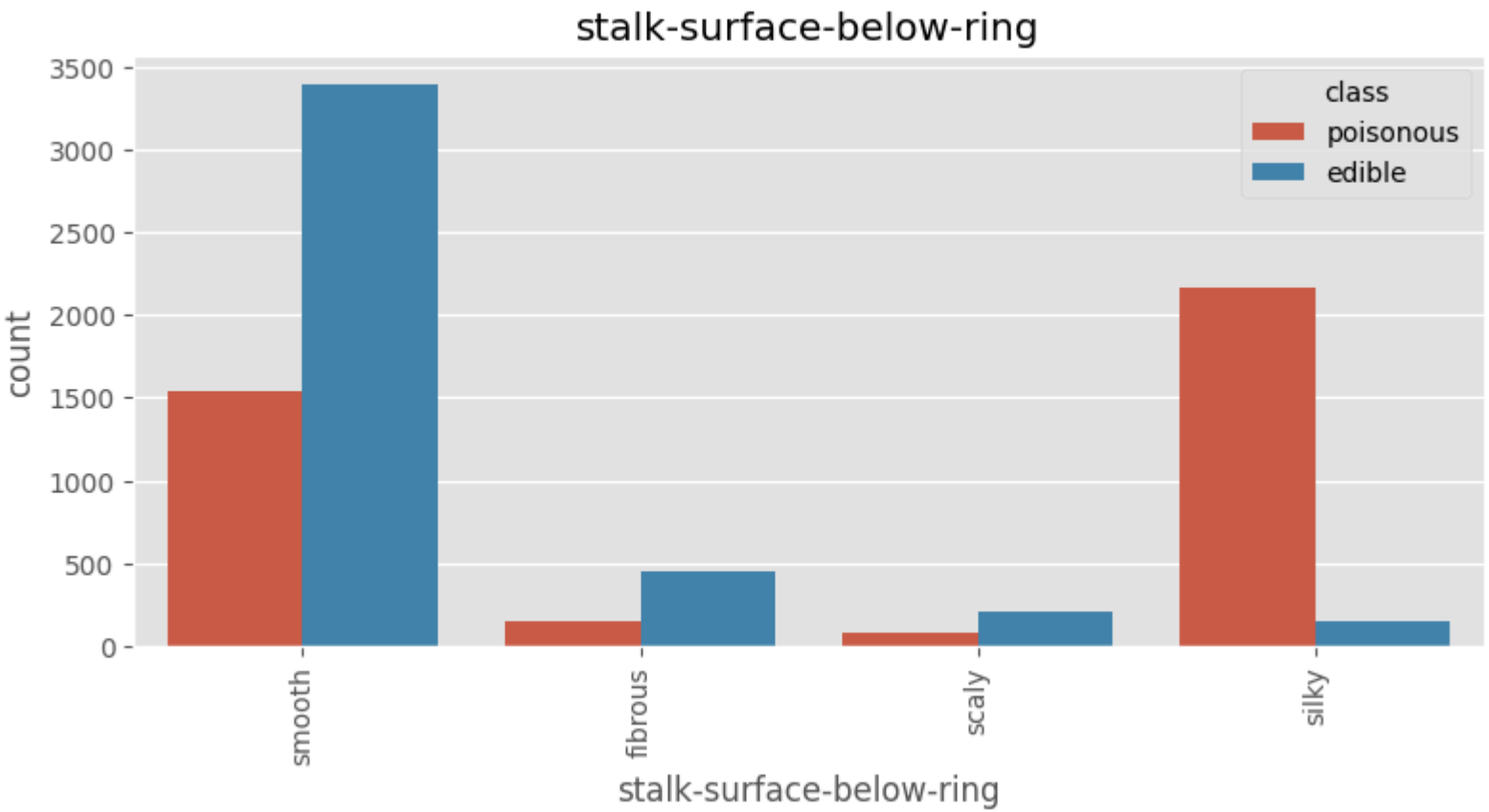

gill-size
class
poisonous
edible
count
4000
3500
3000
2500
2000
1500
1000
500
0
narrow
broad
gill-size

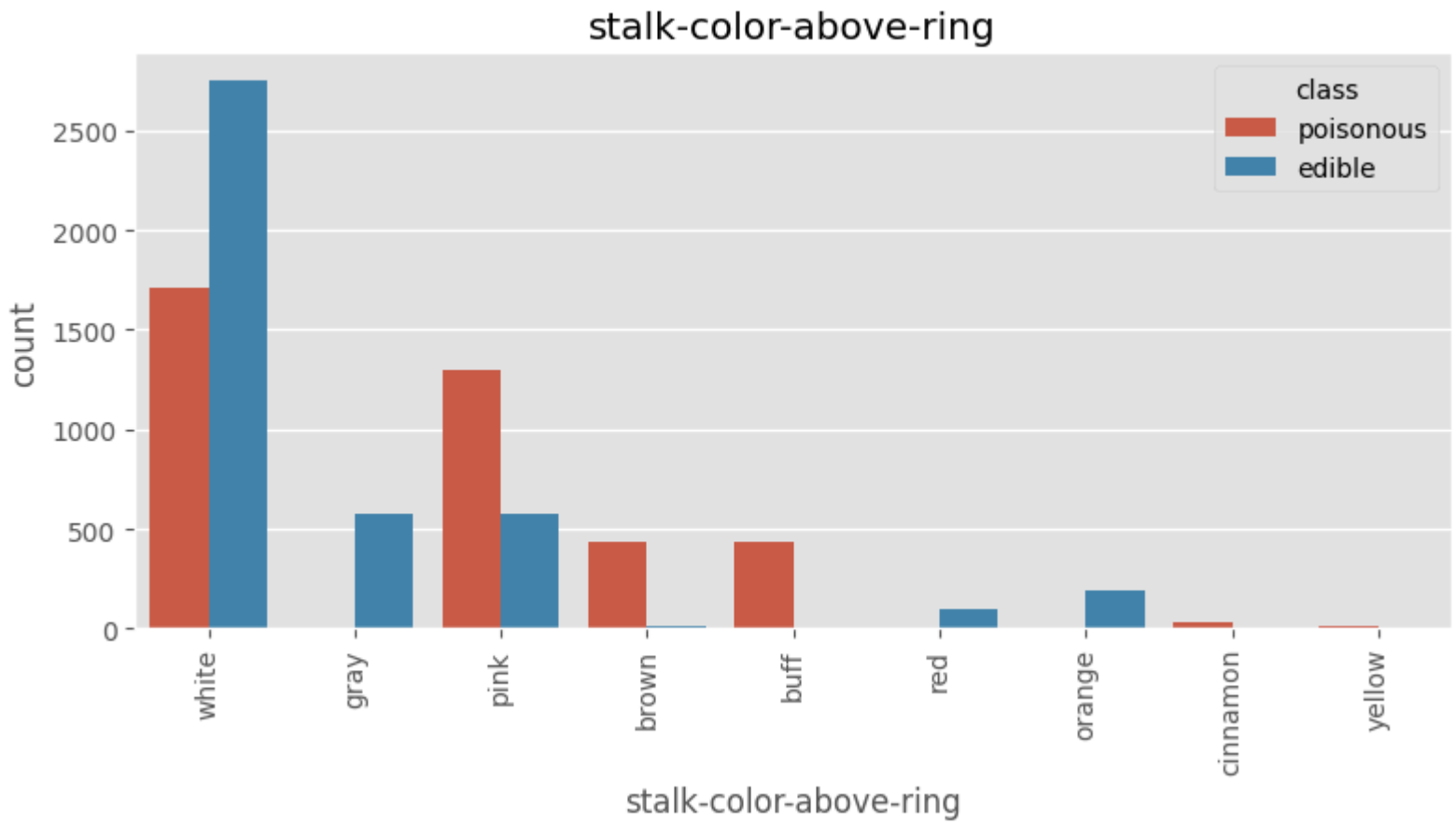

gill-color
class
poisonous
edible
count
1750
1500
1250
1000
750
500
250
0
black
brown
gray
pink
white
chocolate
purple
red
buff
green
yellow
orange
gill-color

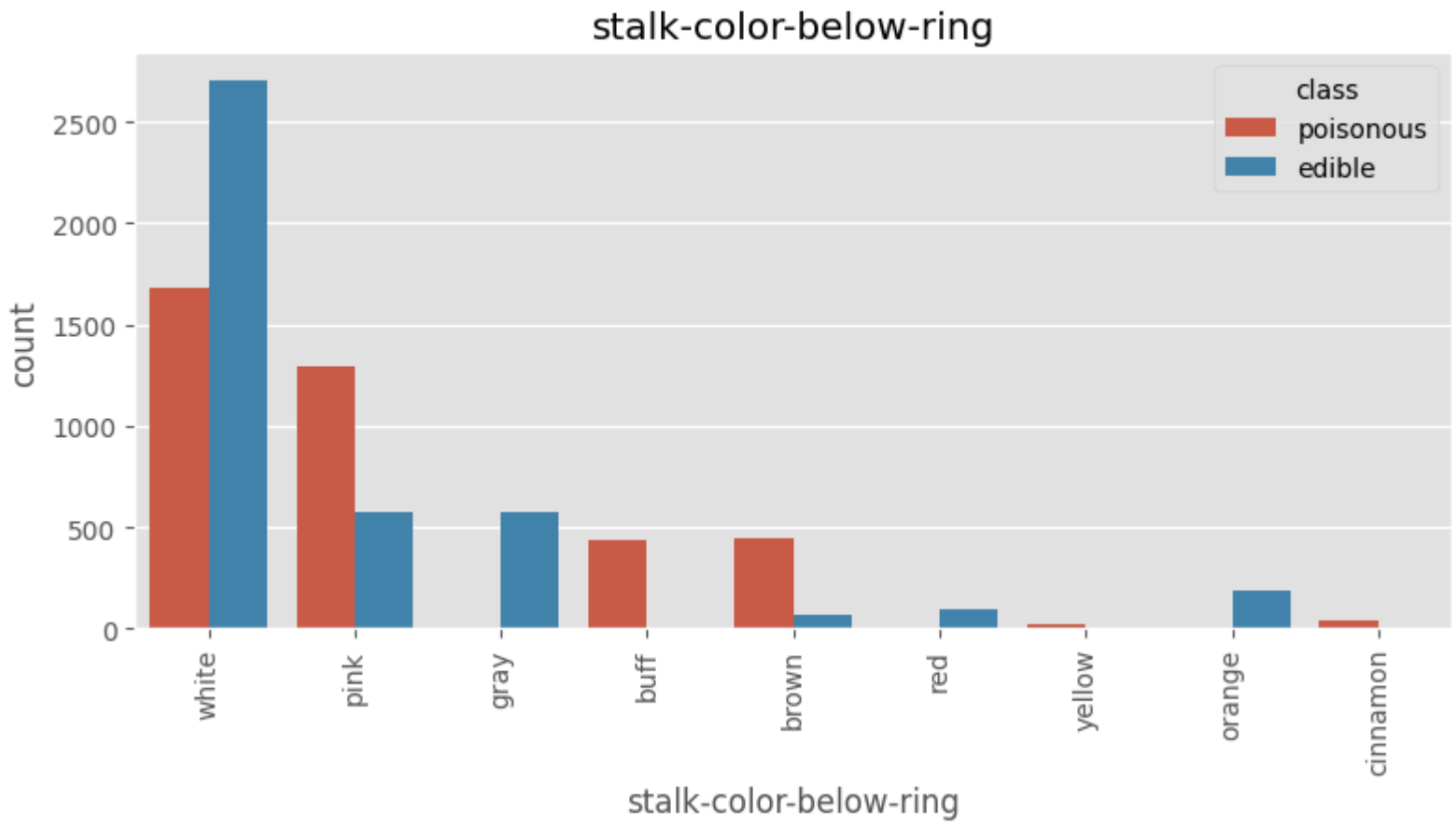

stalk-shape
class
poisonous
edible
2500
2000
1500
1000
500
0
count
enlarging
tapering
stalk-shape

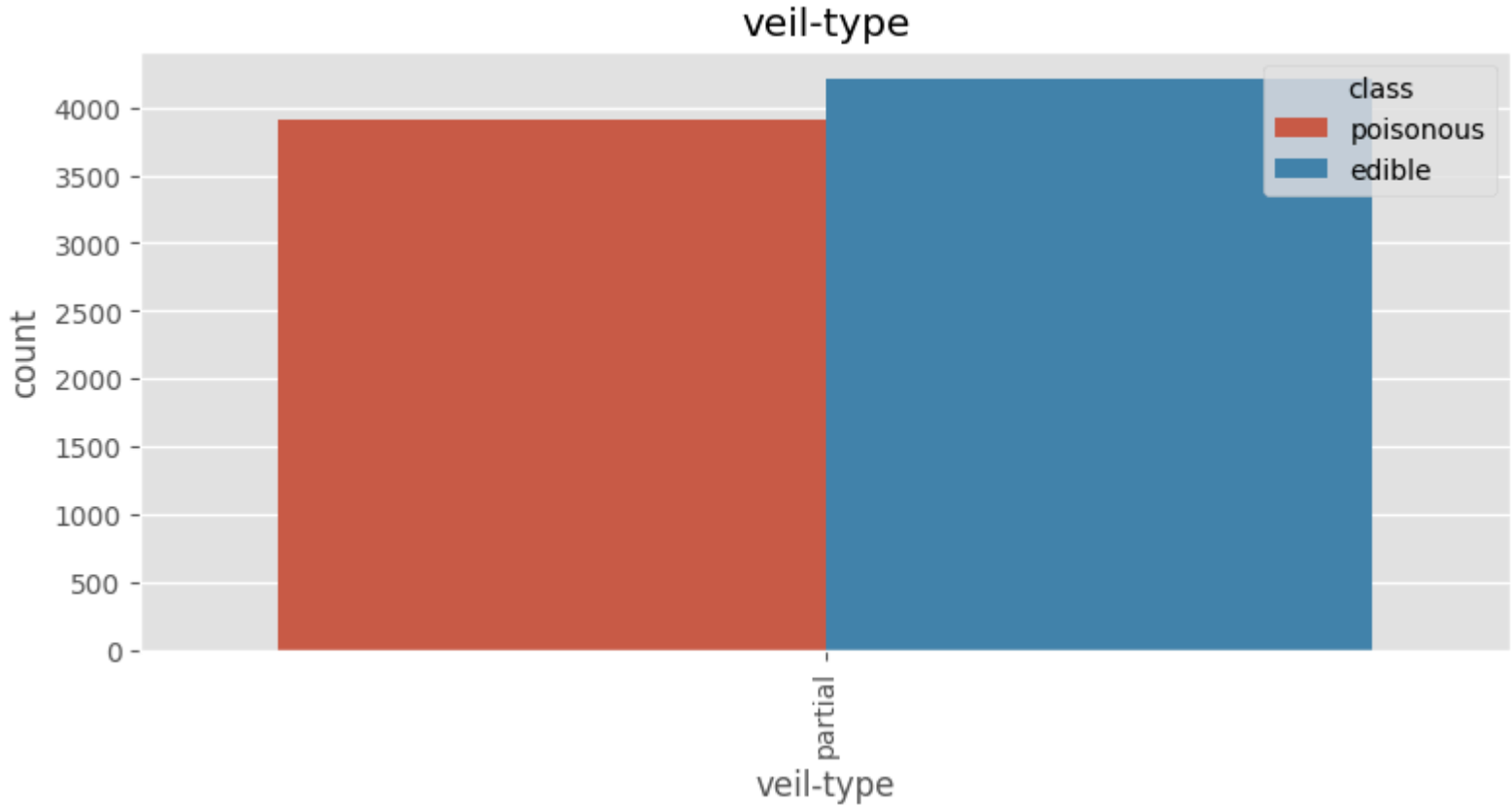

stalk-root
class
poisonous
edible
count
2000
1750
1500
1250
1000
750
500
250
0
equal
club
bulbous
rooted
missing
stalk-root

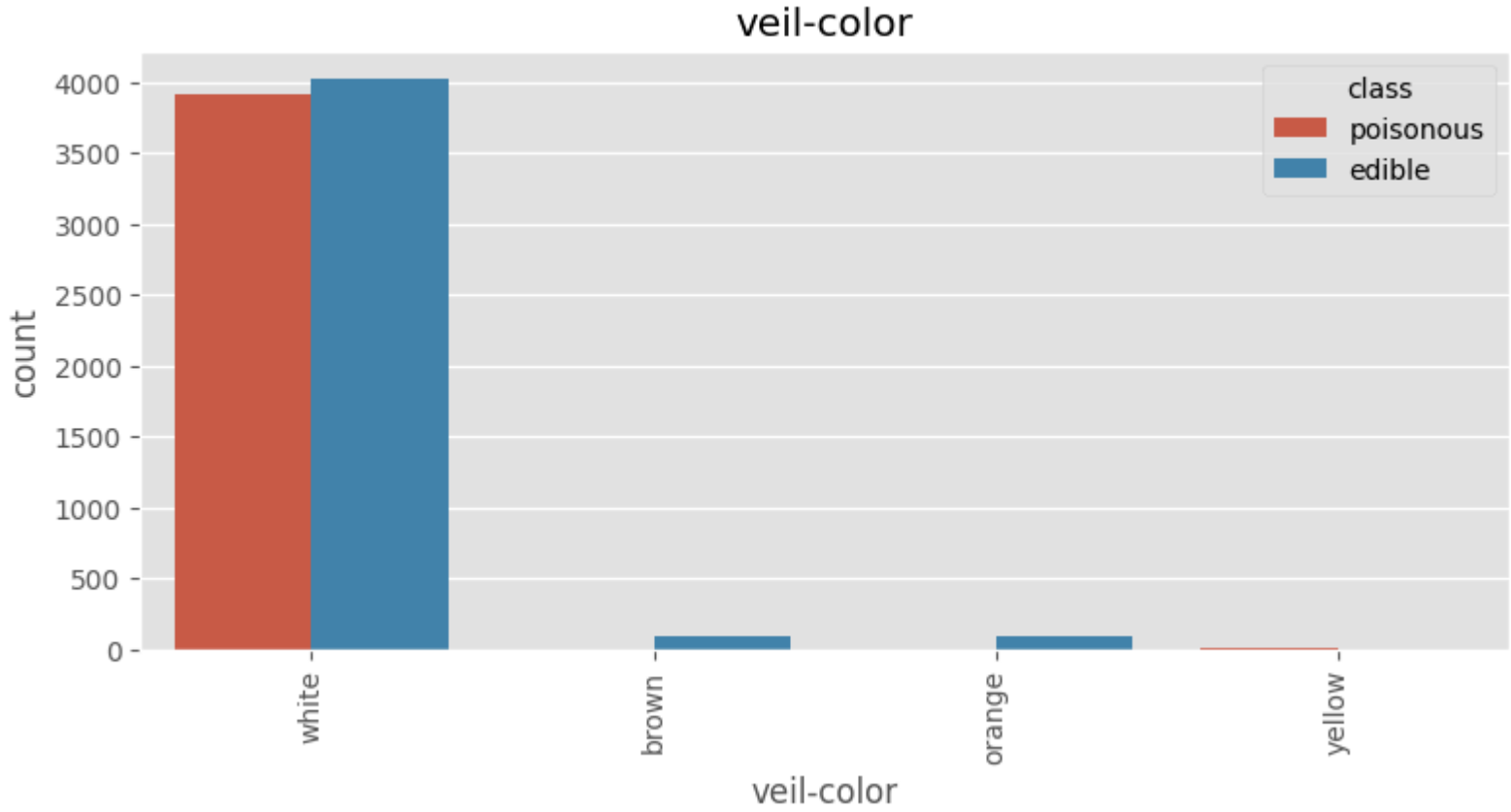

stalk-surface-above-ring
class
poisonous
edible
count
3500
3000
2500
2000
1500
1000
500
0
smooth
fibrous
silky
scaly
stalk-surface-above-ring

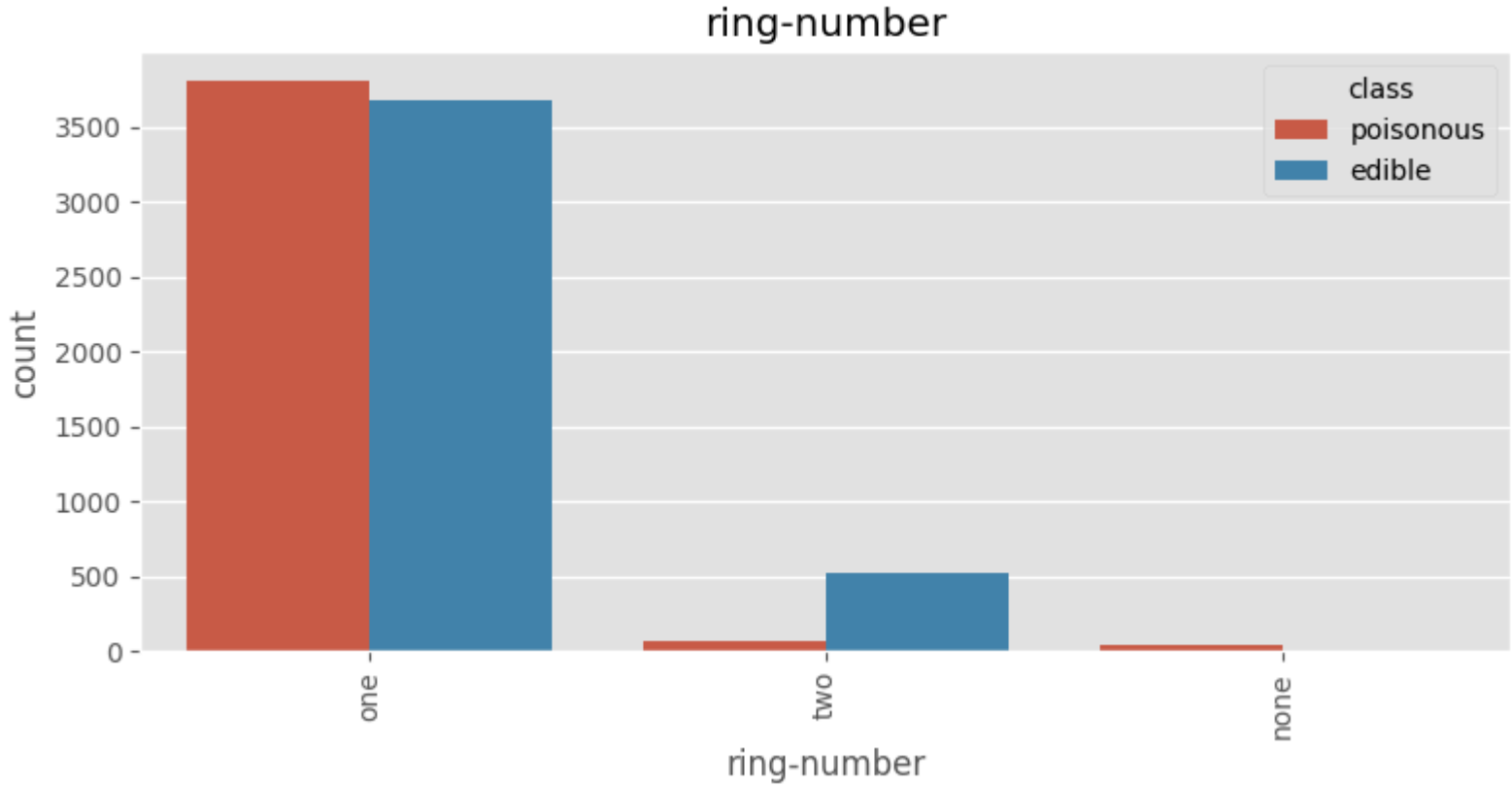

stalk-surface-below-ring
class
poisonous
edible
count
3500
3000
2500
2000
1500
1000
500
0
smooth
fibrous
scaly
silky
stalk-surface-below-ring

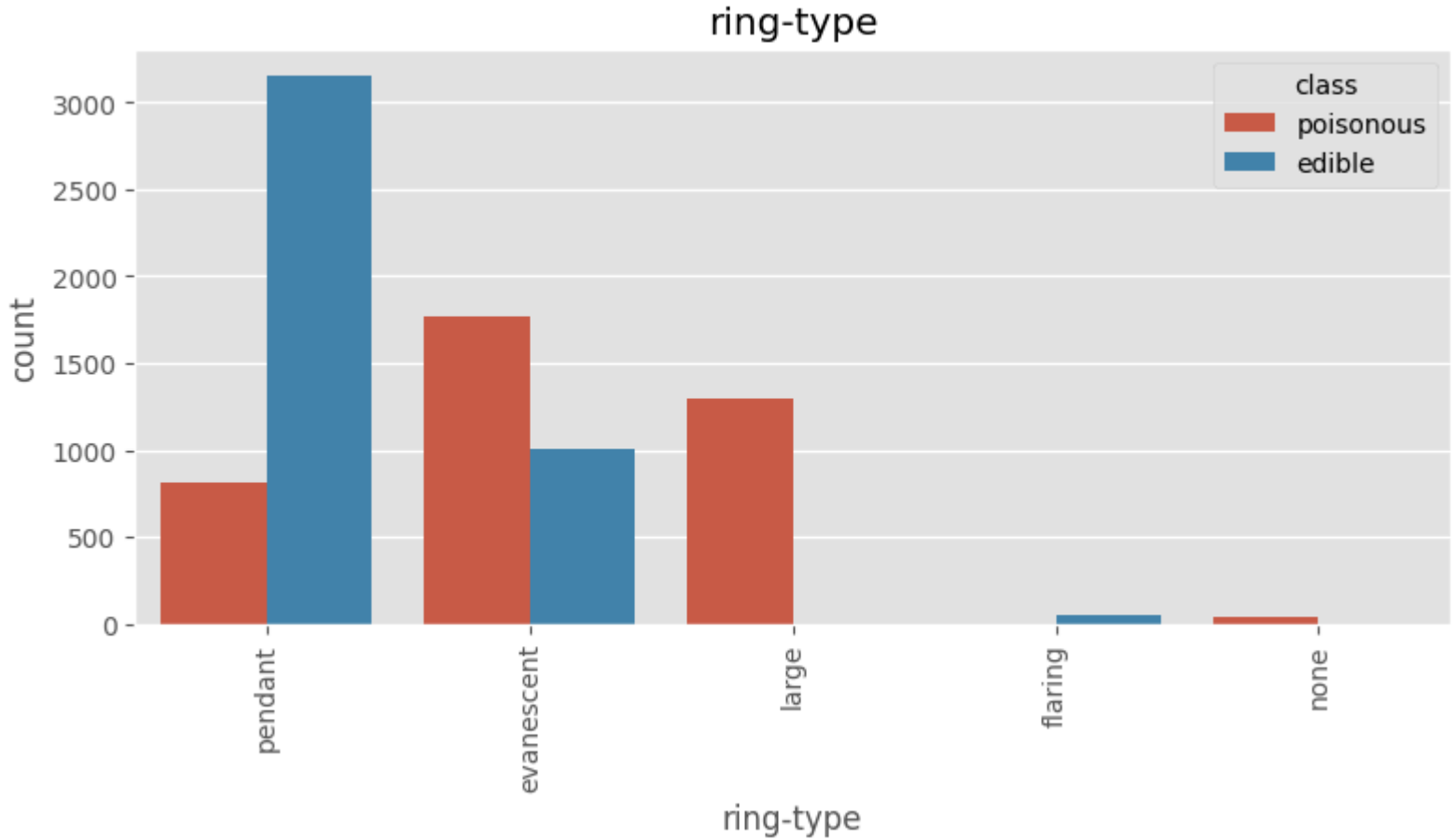

stalk-color-above-ring
class
poisonous
edible
count
2500
2000
1500
1000
500
0
white
gray
pink
brown
buff
red
orange
cinnamon
yellow
stalk-color-above-ring

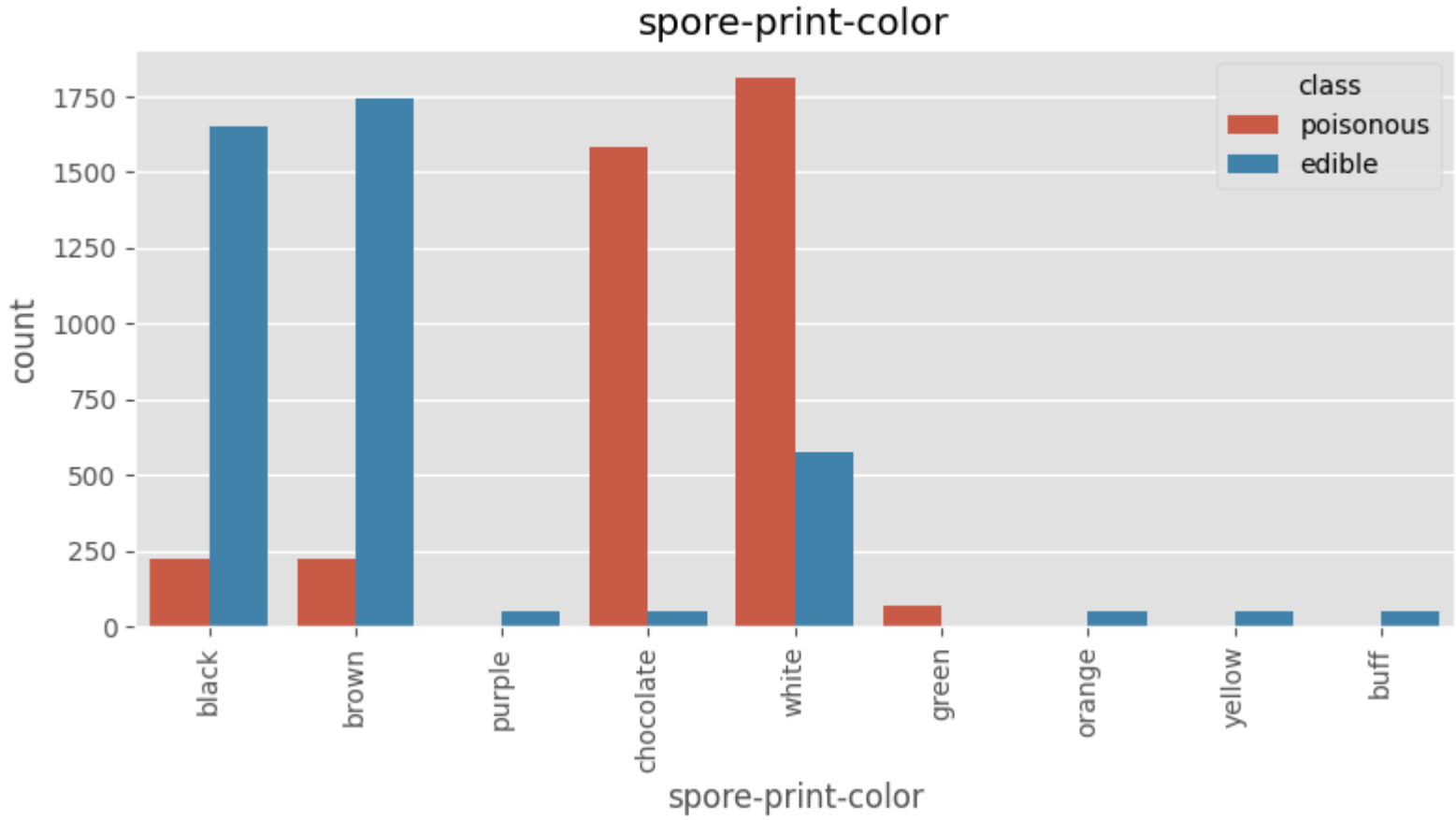

stalk-color-below-ring
class
poisonous
edible
count
2500
2000
1500
1000
500
0
white
pink
gray
buff
brown
red
yellow
orange
cinnamon
stalk-color-below-ring

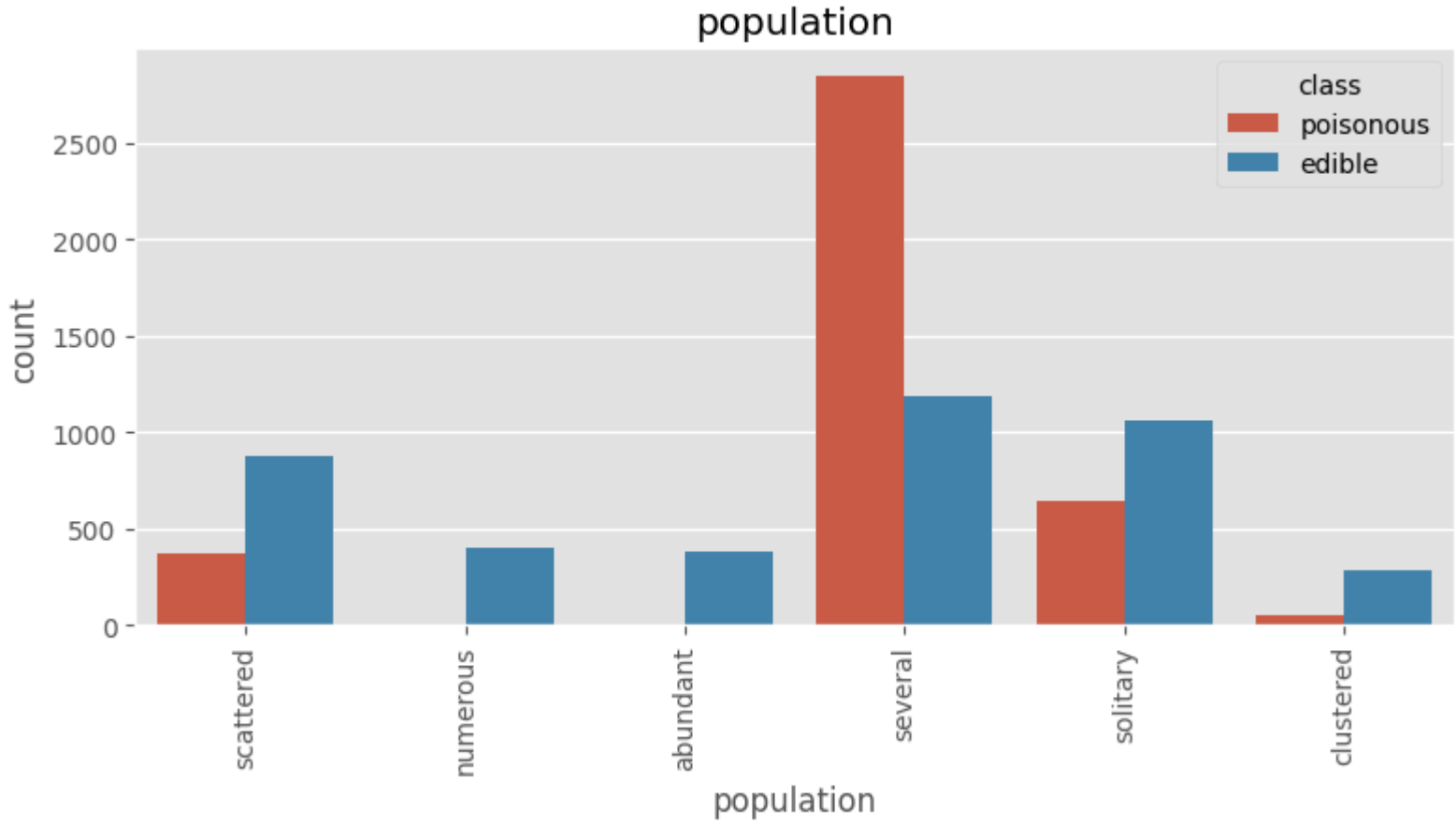

veil-type
class
poisonous
edible
count
4000
3500
3000
2500
2000
1500
1000
500
0
partial
veil-type

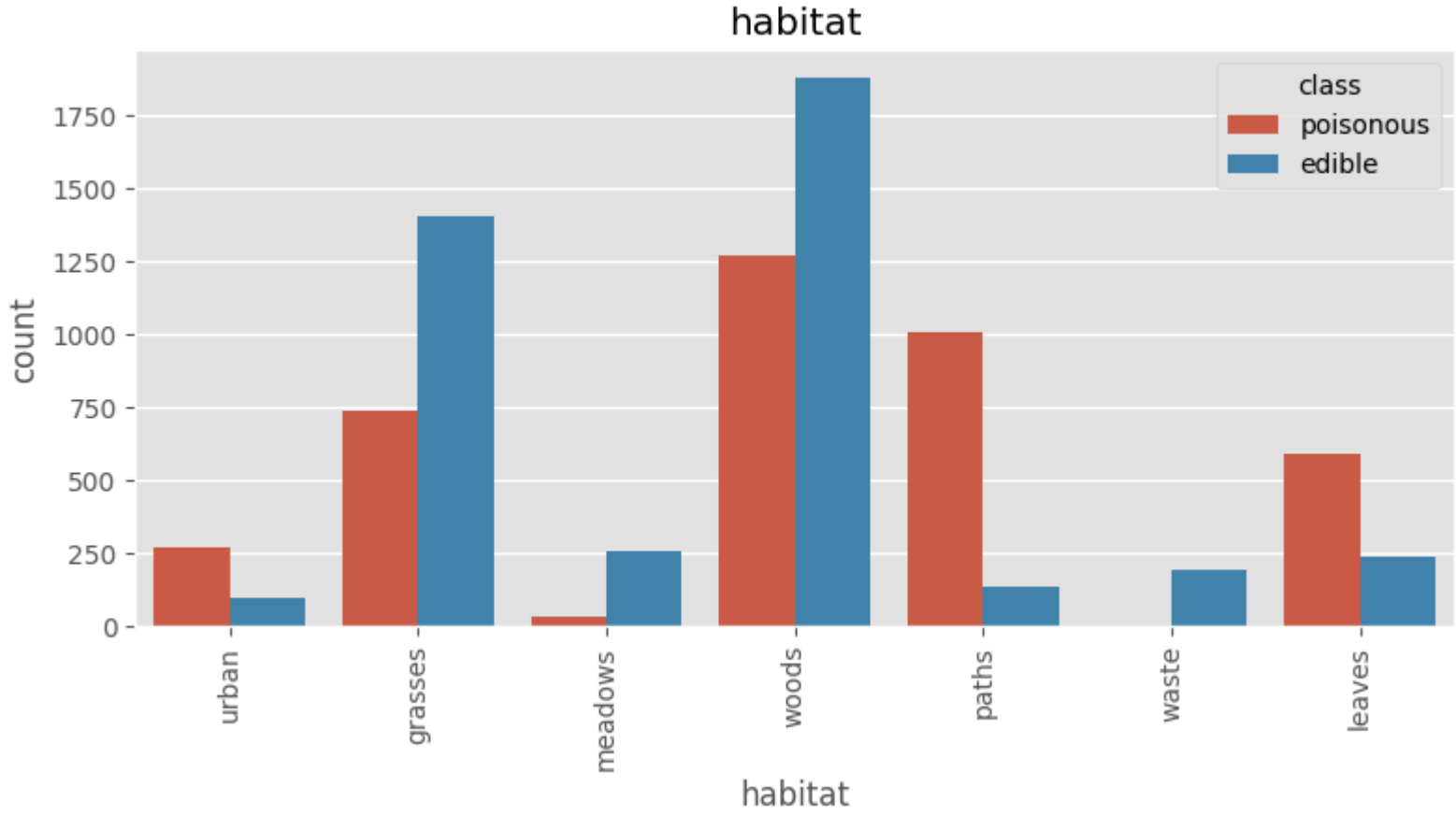

veil-color
class
poisonous
edible
count
4000
3500
3000
2500
2000
1500
1000
500
0
white
brown
orange
yellow
veil-color


ring-number
class
poisonous
edible
count
3500
3000
2500
2000
1500
1000
500
0
one
two
none
ring-number


ring-type
class
poisonous
edible
count
3000
2500
2000
1500
1000
500
0
pendant
evanescent
large
flaring
none
ring-type


spore-print-color
class
poisonous
edible
count
1750
1500
1250
1000
750
500
250
0
black
brown
purple
chocolate
white
green
orange
yellow
buff
spore-print-color


population
class
poisonous
edible
count
2500
2000
1500
1000
500
0
scattered
numerous
abundant
several
solitary
clustered
population


habitat
class
poisonous
edible
count
1750
1500
1250
1000
750
500
250
0
urban
grasses
meadows
woods
paths
waste
leaves
habitat